計測自動制御学会東北支部 第 357 回研究集会 (2026.3.11)
資料番号 357-8

学習済み画像トークナイザによる音響信号の解釈可能性に関する基礎的検討

A Basic Study on Interpretability of Acoustic Signals Using Pretrained Image Tokenizers

○岡部豪太 *，田中大介 *，張山昌論 *

○ Gota Okabe*, Daisuke Tanaka*, Masanori Hariyama*

*東北大学

*Tohoku University.

キーワード： 画像トークナイザ (image tokenizer), 解釈可能性 (interpretability), 説明可能 AI (explainable artificial intelligence), 特徴抽出 (feature extraction), Vision Transformer (vision transformer)

連絡先： 〒 980-8579　仙台市青葉区荒巻字青葉 6-3-09 張山・ウィッデヤスーリヤ研究室
岡部豪太, Tel.: (022)795-7155 E-mail: okabe.gota.q1@dc.tohoku.ac.jp

1. はじめに

音は人間にとって最も基本的かつ普遍的な情報源の一つである. 我々は聴覚を経由し周囲の環境を理解し, 重要な状況判断を下す. 工学分野では, 音響信号は様々な領域における異常検知のための重要な情報源として利用されている. 例えば, 音響共鳴による缶詰の欠陥検知 [1], 呼吸音による肺疾患の診断 [2], そしてトンネル天井パネルの打音試験による異常箇所の特定といった構造健全性の検査 [3] などが挙げられる.

情報工学の発展に伴い, 音響異常検知のための機械学習技術は大きく進化した. 従来の手法は, 主にメル周波数ケプストラム係数（MFCC）などの人為的に作成された特徴量に依存していた. 近年では, 深層学習へのパラダイムシフトが起こり, さらに高精度な結果が得られるようになっている. 例えばメルスペクトログラムを入力とした畳み込みニューラルネットワーク（CNN）は既存の MFCC ベースのアプローチよりも高精度に音声感情認識が可能であることが報告されている [4]. また WavTokenizer[5] は, 高度なニューラルオーディオコーデックにより複雑な音響データの表現を可能としている.

これらの予測精度の目覚ましい向上にもかかわらず, その解釈可能性に焦点を当てた研究は依然として限られている.audioLIME[6] はモデル出力を説明するために提案されているが, その手法は,AudioMNIST[7] のようなベンチマークデータセットとともに,CNN ベースのアーキテクチャに焦点を当てている. したがって, これらのアプローチは,Transformer ベースのモデルや離散表現学習における最近の進歩を十分に反映できていない.

音響信号処理がより複雑なブラックボックスモデルへと移行するにつれて, 性能を犠牲にす

ることなくその出力を解釈する手法が求められている. 本研究では, 解釈性の高い離散画像トークナイザの音響領域への適用可能性を探求し, このギャップを埋めることを目指す.

2. 背景と準備

2.1 ベクトル量子化 (VQ)

ベクトル量子化 (VQ) は, 連続空間のベクトル $\mathbf{e} \in \mathbb{R}^d$ をコードブックと呼ばれる代表ベクトル $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^N$ に写像する処理である. ここで, 各 $\mathbf{c}_i \in \mathbb{R}^d$ はコードワードと呼ばれる. 写像 $q(\cdot)$ は, 以下のように最も近いコードワードのインデックスを選択することで定義される.

$$q(\mathbf{e}) = \arg\min_i \|\mathbf{e} - \mathbf{c}_i\|_2 \tag{1}$$

コードブック $\mathcal{C}$ の学習には複数の手法が存在する.VQ-GAN[8] や TiTok[9] のようなアーキテクチャでは, モデルが再構成誤差を最小化する過程でエンドツーエンドで学習される.

2.2 VQ-GAN

VQ-GAN[8] は,Transformer モデルの潜在表現をベクトル量子化することで画像を二次元の整数列で表現する.

まず入力画像 $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ は畳み込み層などで構成されたエンコーダにより特徴抽出され, 潜在表現 $\mathbf{L}_{2D} \in \mathbb{R}^{\frac{H}{f}\times\frac{W}{f}\times D}$ に変換される. ただし,f はダウンサンプリング係数である.

次に,$\mathbf{L}_{2D}$ の各要素 $l \in \mathbb{R}^D$ は, コードブック $\mathcal{C}$ によりベクトル量子化され,2 次元の離散トークン $\mathbf{Z}_{2D} \in \mathbb{R}^{\frac{H}{f}\times\frac{H}{f}}$ が生成される.

デコーダ $\mathcal{D}$ は,2 次元離散トークンからに対応するコードブックを参照し, 画像 $\hat{\mathbf{I}} \approx \mathbf{I}$ を再構成する.

2.3 TiTok

VQ-GAN は 2 次元の空間的な対応関係を維持するのに対し,TiTok[9] は K 個の要素からなる 1 次元のトークン列に圧縮することで, 表現効率を向上させる.

1) **埋め込み:** 画像 $\mathbf{I}$ はパッチに平坦化され, 埋め込み表現 $\mathbf{X} \in \mathbb{R}^{\frac{H}{f}\times\frac{W}{f}\times d}$ に線形射影される.
2) **潜在トークン表現の連結:** K 個の学習可能な潜在トークン表現 $\mathbf{T} \in \mathbb{R}^{K\times d}$ が埋め込み表現 $\mathbf{X}$ に連結される.
3) **ViT ベースのエンコーディング:** Vision Transformer (ViT) のデコーダにより連結された潜在表現から特徴抽出を行い, グローバル特徴を K 個の潜在トークンに集約する.
4) **1 次元量子化:** VQ は K 個のトークンにのみ適用され, コンパクトな 1 次元離散表現 $\mathbf{z} \in \{1,\ldots,N\}^K$ を生成する.

2.4 One-D-Piece

前節の TiTok には, トークン長が固定であり, 動的な調整ができないという課題がある. この問題を解消するため,One-D-Piece[10] は,Tail Token Drop と呼ばれる正規化手法を提案した.Tail Token Drop を導入したモデルでは, トレーニング中,K 個のトークンの末尾がランダムに切り捨てられる. これにより, エンコーダーは重要な情報を先頭トークンへ集約するように学習され, 各トークンが重要度に応じた役割を担うようになる. Tail Token Drop の導入はトークン化された画像の解釈性向上に寄与する. 本手法導入により,One-D-Piece では再構成精度とトークン長の間で柔軟なトレードオフが可能になり, トークン列 $\mathbf{z}$ 内の個々のトークンが, 異なるレベルの意味的または構造的詳細を表すことが報告されている.

3. 提案手法：離散画像トークナイザによる音響信号解析

3.1 スペクトログラムによるトークン化パイプライン

本研究では, 離散画像トークナイザを用いて音響信号を解析するために, スペクトログラムに変換し, それをトークン化して解析するというアプローチをとる.

1) **短時間周波数変換（STFT）：** 生の音響信号は, 短時間フーリエ変換（STFT）を用いてスペクトログラムに変換される. 本実験で利用した画像トークナイザはカラー画像以外を処理できないため, スペクトログラムを 3 チャンネル画像 $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ に変換して扱う.

2) **離散トークン化：** 周波数分解能を確保するため, スペクトログラムは周波数方向に 2 分割して事前学習済みの離散画像トークナイザ（TiTok もしくは One-D-Piece）に入力される. それぞれのスペクトログラムは K 個の離散トークンからなる 1 次元シーケンス $\mathbf{z} \in \{1,\ldots,N\}^K$ に圧縮される. 本研究では,$K = 64$ トークンをそれぞれのスペクトログラムから得る.

3) **線形分類：** 抽出された 2 画像分のトークン列を連結し, 線形分類器を学習させて, 離散表現から元音声がどのクラスに属すのかを予測する.

3.2 SHAP と累積寄与率による解釈可能性分析

特定のトークンが分類タスクにおいて支配的な役割を果たしているかどうかを評価するために,SHAP (SHapley Additive exPlanations) [11] を利用する. 表現の解釈可能性は, 各トークンの寄与度に基づいて定義される. モデルの解釈可能性が高い場合, 一部のトークンがタスク関連情報の大部分を表現しているはずである. したがって, トークンごとの寄与度の偏りを測定することで, 解釈可能性を評価できると考えられる. 以上の仮定に基づき, 本研究では解釈可能性を累積寄与率により評価する. まず分類タスクについて線形モデルによる One-vs-Rest (OvR) 分類器を学習し, 予測に対するトークンごとの SHAP 値を算出する. 次に全クラスの SHAP 値の絶対値を集計し, 累積寄与率曲線をプロットすることで, 情報の分布を視覚化する. 大きく上向きに湾曲した曲線は, 特定のトークンがタスクに対して不均衡に大きな貢献をしていることを示す. Tail Token Drop はトークン列の先頭側が画像全体の特徴を表現するようになる正則化手法である. したがって Tail Token Drop が解釈性に寄与する場合, スペクトログラムそのものの特徴を表現する先頭側のトークンは, 分類タスクに対して寄与が小さくなり, 細かい差異を表現する後ろ側のトークンの寄与が大きくなると予想される.

4. 実験と評価

4.1 実験設定

前章で述べた処理パイプラインにより,Tail-Token Drop を採用した One-D-Piece と TiTok を比較して, 音響信号の解釈可能性を向上させるかを検討した. 実験では,6 人の話者が数字 0〜9 を発音した録音（話者 1 人につき各数字 50 サンプル）からなる Free-Spoken-Digit-Dataset (FSDD) を使用し, 話者識別をタスクとして設定した. なお, 録音環境やマイクゲインの影響を排除するため, すべての音響信号に対して前処理として,STFT を行う前に最大音量へとピーク正規化を行った.

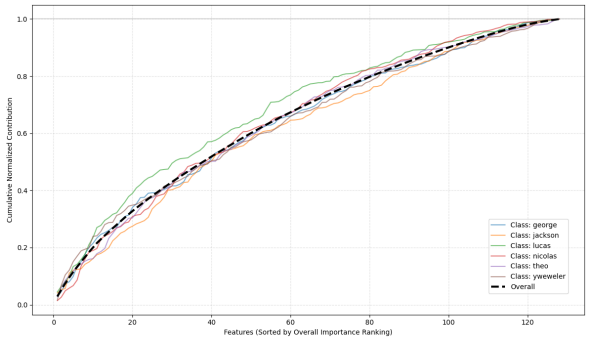

Fig. 1　One-D-Piece の累積寄与率

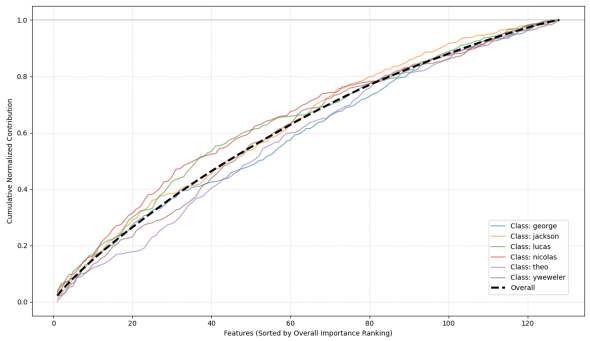

Fig. 2　TiTok の累積寄与率

4.2　結果

図 1 に示される One-D-Piece の累積寄与率は図 2 に示された TiTok の累積寄与率よりもわずかに上方に偏った, 凸性の強い累積寄与率を示した. 上位 40 特徴量について累積寄与率を比較すると TiTok では 0.47 程度なのに対して,One-D-Piece では,0.52 程度となった. この結果は,One-D-Piece に導入された Tail Token Drop により, 一部のトークンがタスクに特化した特徴を集中的に表現している可能性を裏付けている. しかし, クラスごとに最も寄与しているトークンは One-D-Piece, TiTok のいずれも分散した. 図 3 と図 4 はそれぞれ One-D-Piece と TiTok のクラスごとの SHAP 値の合計を昇順に示しており, 縦軸のラベルは分割された画像のうち高音域（H）と低音域（L）のどちらのスペクトログラムなのかを, 数字は各画像のトークン列中でのインデックスを表している. これらのグラフはいずれも分類されるクラスによって寄与が最も大きいトークンが異なり, タスクに特化したトークン, すなわち話者性などのタスク解析に必要な情報を保持する役割のトークンが存在しないことを示唆している. また, 図 3 は高音域側の先頭トークン（H_1, H_2）の平均寄与度が大きいことも示しており,Tail Token Drop の性質から予想される事象とは異なり分割したスペクトログラム画像に対するトークン列の先頭側にはスペクトログラムそのものを表現するトークンが含まれておらず, 反対にクラス分類に必要なスペクトログラムの違いを表すトークンが含まれていると考えられる. この理由については次章で考察する.

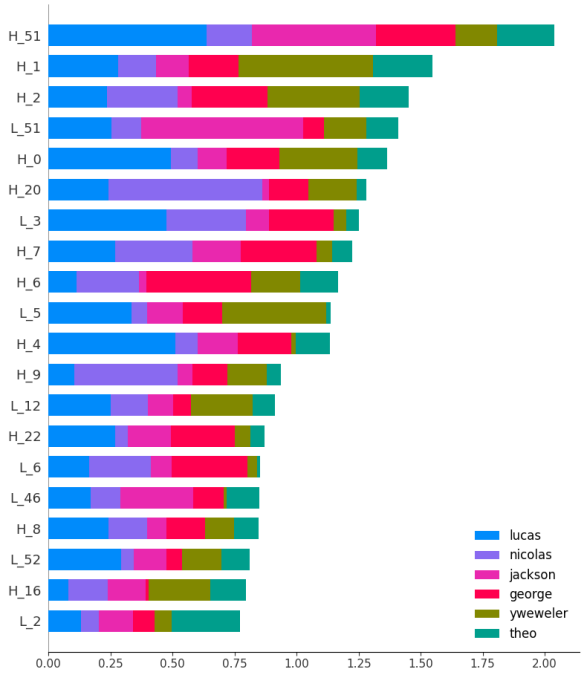

Fig. 3　One-D-Piece の SHAP 値

5.　おわりに

本研究では, スペクトログラムに対して離散画像トークナイザ, 具体的には TiTok と One-D-Piece を音響信号解析に適用した場合の解釈可能性を検討した. 累積寄与率を分析した結果,One-D-Piece は TiTok と比較して特徴量の重要度がある程度集中していることが明らかになった. これは,Tail Token Drop メカニズムがスペクトログラムに対しても有効に機能していることを示

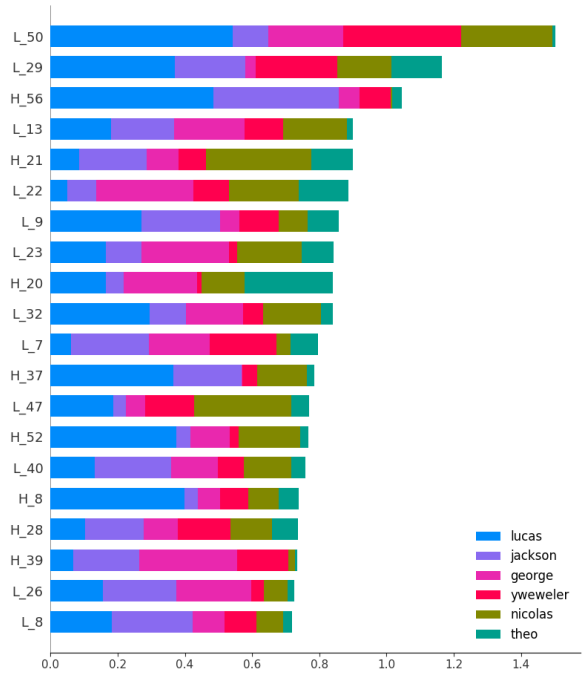

Fig. 4　TiTok の SHAP 値

唆している．一方でクラスごとに SHAP により分析を行った結果, 各クラスの分類に最も大きく寄与するトークンは一貫性がなく, 異なる位置のトークンに散在していることが示された. この一貫性の欠如が示すのは, 現状の学習済みモデルをそのまま適用するだけでは特定のトークンを話者識別などのタスク関連特徴の普遍的な指標として確立できていない可能性である.Tail Token Drop から予想される結果と反してスペクトログラムそのものを表現するトークンが存在しない可能性も示されており, 音響領域において個々のトークンが果たす具体的な意味的または構造的役割を, 本実験結果からは明確に特定することは困難であった.

これらの限界は, 主に 2 つの要因に起因すると考えられる. 第一に, 空間処理における根本的な不一致である．TiTok や One-D-Piece のような 1D トークナイザは, 厳密な位置依存性なしに画像全体の情報を取得するように設計されているが, 本アプローチでは周波数分解能を維持するためにスペクトログラムを分割した. この断片化により, トークナイザが音響情報をまとまりのある全体的表現として処理する能力が損なわれた可能性がある.

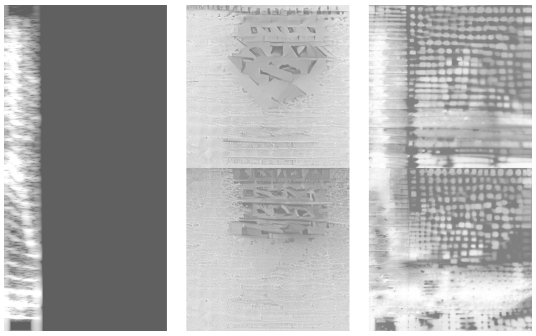

Fig. 5　元のスペクトログラム（左）と TiTok（中）, One-D-Piece（右）により復元されたスペクトログラム

第二に, 本研究で用いたトークナイザがスペクトログラムの解析に適応できていなかった可能性も考えられる. 図 5 は元のスペクトログラムを TiTok と One-D-Piece によりそれぞれ復元したものである. 復元結果はいずれも元のスペクトログラムの構造を完全に喪失しており, 適切に復元てきているとはいえない. これは, 自然画像で学習されたトークナイザが, スペクトログラム固有の構造と特徴を十分に処理できなかったことを示している.

以上から本研究で提案したアーキテクチャは Tail Token Drop が音響信号の解釈性向上に有意義である可能性を示唆したものの, 具体的な解釈性向上までを立証するには至らなかった.

今後は音響信号解析に特化して最適化された専用のトークナイザを開発することで, これらの障壁を克服することに焦点を当てる. これには, フルパラメータ更新を用いたモデルの微調整, あるいは大規模な音響データセットを用いて専用のアーキテクチャをゼロから学習することが含まれる. 潜在表現を音声データのニュアンスに合わせて調整することで, 複雑な音響タスクにおいて, より堅牢で解釈可能な離散表現を獲得することを目指す.

参考文献

1) 谷川英一, 元広輝重, 秋場稔 編著. 缶詰製造学, 恒星社厚生閣, (1969)

2) Sandra Reichert, Raymond Gass, Christian Brandt, Emmanuel Andrès: Analysis of respiratory sounds: state of the art." Clinical medicine. Circulatory, respiratory and pulmonary medicine 2, pp. 45–58, (2008)

3) トンネル天井板の落下事故に関する調査・検討委員会: 「トンネル天井板の落下事故に関する調査・検討委員会」資料集, 395/415, 国土交通省 (2013)

4) Bonaventure F. P. Dossou and Yeno K. S. Gbenou: FSER: Deep Convolutional Neural Networks for Speech Emotion Recognition, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pp. 3533–3538, (2021)

5) Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng and Zhou Zhao: WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling, arXiv preprint arXiv:2408.16532, (2024)

6) Verena Haunschmid, Ethan Manilow, Gerhard Widmer: audioLIME: Listenable Explanations Using Source Separation, arXiv preprint arXiv:2008.00582, (2020)

7) Sören Becker and Johanna Vielhaben and Marcel Ackermann and Klaus-Robert Müller and Sebastian Lapuschkin and Wojciech Samek: AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark, Journal of the Franklin Institute, (2023)

8) Patrick Esser, Robin Rombach, Björn Ommer: Taming Transformers for High-Resolution Image Synthesis, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12873–12883, (2021)

9) Qihang Yu, Mark Weber, Xueqing Deng, Xiaohui Shen, Daniel Cremers and Liang-Chieh Chen: An Image is Worth 32 Tokens for Reconstruction and Generation, Advances in Neural Information Processing Systems 37, pp. 128940–128966, (2024)

10) K. Miwa, K. Sasaki, H. Arai, T. Takahashi and Y. Yamaguchi: One-D-Piece: Image Tokenizer Meets Quality-Controllable Compression, Tokenization Workshop (2026)

11) Scott M. Lundberg and Su-In Lee: A unified approach to interpreting model predictions, Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777, (2017)