計測自動制御学会東北支部 第 356 回研究集会 (2025.12.10)
資料番号 356-7

音声トークナイザを用いた異常検知手法の基礎的検討 —機械音データへの適用評価—

A Exploratory Study on Anomaly Detection Methods Using an Audio Tokenizer: Evaluation on Machine Sound Data

○田中 大介*，岡部 豪太*，張山 昌論*

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

*東北大学

*Tohoku University

キーワード： WavTokenizer，教師なし異常検知 (Unsupervised Anomaly Detection)，
ニューラル音響コーデック (Neural Audio Codec)，音響言語モデル (Audio Language Model)，
ドメインシフト (Domain Shift)

連絡先： 〒 980-8579 仙台市青葉区荒巻字青葉 6-3-09
東北大学 大学院情報科学研究科 張山・ウィッデヤスーリヤ研究室 田中大介，
Tel.: (022)795-7155，E-mail: daisuke.tanaka.b7@tohoku.ac.jp

1. はじめに

近年，音響信号を離散的なトークン列として表現する音声トークナイザ（Audio Tokenizer）が，音声生成や高効率コーデックの分野において基盤技術となりつつある．VQ-VAE に基づく離散表現学習 [1] をはじめ，SoundStream[2] や EnCodec[3] に代表されるニューラル音響コーデック，さらに AudioLM[4] などの音響言語モデルでは，連続的な音響信号を有限の整数トークンへと量子化し，その系列構造をモデル化することで高品質な復元・生成を実現している．特に，近年提案された WavTokenizer[5] は，大規模なコードブックと最適化されたデコーダ構造を採用することで，極めて少ない量子化層数でも従来の多層量子化モデルに匹敵する再構成品質を達成しており，次世代の音響離散表現として注目されている．

これらの音声トークナイザは，高次元で冗長な音響信号を，意味的な構造を保ちつつコンパクトな離散表現へ圧縮するという性質を持つ．この特性は，音声のみならず環境音，音楽，機械音といった多様な音響データに対して，特徴抽出や類似度計算などの解析タスクにおいても有用であると考えられる．特に，整数トークン列という表現形式は，従来の連続値スペクトログラムとは異なる特徴空間を提供する．離散化に伴う情報のボトルネック効果により，正常データの本質的なパターンのみがトークン列として保持され，そこからの逸脱として異常を捉えるという，新たな信号理解の枠組みをもたらす可能性がある．

しかしながら，こうした離散音響表現の応用研究は，現状では音声合成や音楽生成といった生

成タスクが主であり，異常検知（Anomaly Detection）などの識別・解析タスクへの適用に関する知見は限定的である．一般に，異常検知分野ではオートエンコーダ等の再構成誤差に基づく手法が広く研究されているが[6]，離散トークンを用いたアプローチの有効性は十分に検証されていない．

本技術の応用先としては，発話障害の検知や病的音声のスクリーニングといった音声データの異常検知が期待される．しかし，音声データは言語情報・話者性・感情などの多様な変動要因を含んでおり，提案手法（トークン化による異常スコア算出）の有効性を初期段階で検証するには，変動要因が複雑すぎるという課題がある．

そこで本研究では，音声データへの適用に向けた予備的検討として，物理的な正常・異常の定義が明確であり，かつ標準的なベンチマークとして広く利用されている DCASE 2022 Task 2 の機械音データセット[7]を評価対象とする．音声データで学習された WavTokenizer が，ドメインの異なる機械音に対してどの程度の表現能力を持ち，異常検知に寄与できるかを確認することは，手法の基礎特性や汎用性を評価する上で重要なステップであると考えられる．具体的には，WavTokenizer により得られる整数トークン列に着目し，以下の 2 つの異なるアプローチによる異常スコア算出手法について，その有効性と特性を比較検討する．

- **再構成誤差（MSE）に基づく手法:** 離散トークンを経由して復元された音響信号と元信号との誤差を評価する．正常音は適切にトークン化・復元される一方，異常音は量子化誤差が増大するという仮説に基づく．
- **言語モデル（LM）に基づく手法:** トークン列を時系列データとみなし，Transformer 等の言語モデルを用いて「トークンの並び順（文法）」を学習する．正常な機械動作が持つ規則性から逸脱した系列を，尤度の低下として検知する．

本研究では，学習時と同一条件の *source* ドメインと，条件が変動した *target* ドメインの双方において評価を行い，(1) 閉集合条件における異常識別性能，(2) ドメイン変動に対する汎化性能，(3) 定常音や非定常音といった機械特性による有効性の差異，を明らかにする．本稿では，将来的な音響理解システムの構築を見据え，汎用的な音声トークナイザを信号解析タスクへ適用し，その適用可能性に関する基礎的な知見を報告する．

2. WavTokenizerによる音響信号の離散化

本研究では，音響信号の離散表現を得るために，最新のニューラル音響コーデックの一つである WavTokenizer[5] を用いる．本節では，その基本構造と離散トークンの生成過程について述べる．

2.1 ニューラル音響コーデックとベクトル量子化

ニューラル音響コーデックは一般に，エンコーダ（Encoder）$\mathcal{E}$，量子化器（Quantizer）$\mathcal{Q}$，およびジェネレータ（デコーダ） $\mathcal{G}$ の 3 つのモジュールから構成される．

入力される高次元な時系列信号（生波形）を $\mathbf{x} \in \mathbb{R}^T$ とする（T はサンプル数）．まず，エンコーダ $\mathcal{E}$ は $\mathbf{x}$ を低次元の潜在表現 $\mathbf{z}$ へと写像する．

$$\mathbf{z} = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{C \times L} \tag{1}$$

ここで，C はチャネル数，L は時間フレーム数であり，ダウンサンプリングにより $L \ll T$ となる．また，$\mathbf{z}$ は各時刻 t $(1 \leq t \leq L)$ における特徴量ベクトル $\mathbf{z}_t \in \mathbb{R}^C$ を列方向に並べた行

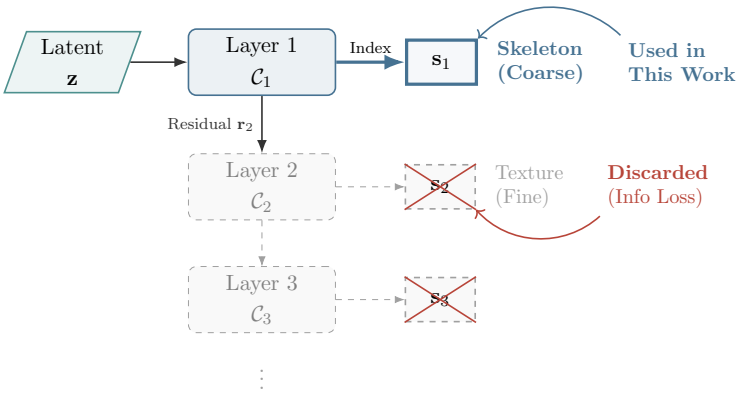

Fig. 1　本研究における WavTokenizer の利用設定

列として表される．

$$\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_L) \tag{2}$$

次に，量子化器 $\mathcal{Q}$ は連続値ベクトル列 $\mathbf{z}$ をフレームごとに離散化する．最も基本的なベクトル量子化（Vector Quantization: VQ）においては，学習可能な単一のコードブック $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_V\}$ を用いる．ここで，各コードベクトルは $\mathbf{c}_i \in \mathbb{R}^C$ であり，V はコードブックサイズ（語彙数）を表す．

量子化の過程では，各時刻 t の入力ベクトル $\mathbf{z}_t$ に対し，コードブックの中からユークリッド距離が最も近いベクトルを探索して置換を行う．

$$s_t = \underset{i \in \{1,\ldots,V\}}{\operatorname{argmin}} \|\mathbf{z}_t - \mathbf{c}_i\|_2 \tag{3}$$

$$\hat{\mathbf{z}}_t = \mathbf{c}_{s_t} \tag{4}$$

ここで，$s_t \in \{1, \ldots, V\}$ は時刻 t におけるトークン ID（スカラー）である．また，これらを時間方向に並べた系列 $\mathbf{s} = (s_1, \ldots, s_L) \in \{1, \ldots, V\}^L$ をトークン列と呼ぶ．

最後に，ジェネレータ $\mathcal{G}$ は量子化された潜在表現 $\hat{\mathbf{z}} = (\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_L)$ から，元の波形を近似した再構成波形 $\hat{\mathbf{x}}$ を生成する．

$$\hat{\mathbf{x}} = \mathcal{G}(\hat{\mathbf{z}}) \approx \mathbf{x} \tag{5}$$

学習においては，エンコーダ $\mathcal{E}$，ジェネレータ $\mathcal{G}$ に加え，量子化器内のコードブック $\mathcal{C}$ も含めた全パラメータが End-to-End で最適化される．具体的には，再構成誤差 $\mathcal{L}_{rec}$ の最小化に加え，連続ベクトル $\mathbf{z}_t$ と離散ベクトル $\hat{\mathbf{z}}_t$ を近づけるためのコードブック損失（Commitment Loss 等）を導入し，入力信号の分布に適合した最適な離散表現を獲得するように学習が行われる．

2.2　残差ベクトル量子化 (RVQ)

前節の単純な VQ では情報の表現能力に限界があるため，WavTokenizer を含む現代的なニューラルコーデックでは，残差ベクトル量子化（Residual Vector Quantization: RVQ）が採用されている．これは，N_q 個の階層的なコードブック $\mathcal{C}_1, \ldots, \mathcal{C}_{N_q}$ を用いて，入力 $\mathbf{z}$ を段階的に近似する手法である．

量子化は再帰的に行われる．第 1 層目の入力（残差）を $\mathbf{r}_1 = \mathbf{z}$ とし，第 k 層目（$1 \le k \le N_q$）の量子化処理は以下のように記述される．まず，現在の残差 $\mathbf{r}_k$ を第 k 層のコードブック $\mathcal{C}_k$ で量子化し，量子化ベクトル列 $\hat{\mathbf{z}}_k$ を得る．次に，次層への入力となる残差を $\mathbf{r}_{k+1} = \mathbf{r}_k - \hat{\mathbf{z}}_k$ として更新する．

最終的な量子化済み潜在表現 $\hat{\mathbf{z}}$ は，全 N_q 層の量子化ベクトル列の和として構成される．

$$\hat{\mathbf{z}} = \sum_{k=1}^{N_q} \hat{\mathbf{z}}_k(\mathbf{s}_k) \tag{6}$$

このとき，各層 k において選択されたトークン

列（インデックスのベクトル）を $\mathbf{s}_k \in \{1, \ldots, V\}^L$ とする．モデル全体が出力する離散表現 $\mathbf{S}$ は，これら N_q 本のトークン列の集合（行列）として定義される．

$$\mathbf{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_{N_q}\} \in \{1, \ldots, V\}^{N_q \times L} \quad (7)$$

2.3 WavTokenizer の特徴と本研究の設定

従来の EnCodec[3] 等が N_q を大きく設定する（例：8 層〜32 層）ことで高音質化を図っていたのに対し，WavTokenizer は，より大きなコードブックサイズ（$V \geq 4096$）と，敵対的生成ネットワーク（GAN）に基づく強力なジェネレータ $\mathcal{G}$ を採用することで，単一の量子化層（$N_q = 1$）であっても従来の多層モデルに匹敵する再構成品質を達成している点に特徴がある．

本研究では，この $N_q = 1$ で動作するという WavTokenizer 固有の特性に着目する．本研究におけるモデルの利用設定と情報の取捨選択の概念図を図 1 に示す．この場合，式 (6) における和は不要となり，第 1 層目の量子化ベクトル列のみが用いられる（$\hat{\mathbf{z}} = \hat{\mathbf{z}}_1$）．同様に，トークン列も $\mathbf{S} = \mathbf{s}_1$ となる．

$$\mathbf{S} = \mathbf{s}_1 \in \{1, \ldots, V\}^L \quad (8)$$

この設定には，異常検知タスクにおいて以下の利点があると期待される．第一に，系列の単純化である．複数の残差層を階層的に扱う必要がなく，単一のトークン列として扱えるため，標準的な言語モデル（Transformer 等）による文脈学習が容易に適用可能である．第二に，情報の集約である．強力なジェネレータ $\mathcal{G}$ による復元を前提としているため，第 1 層のトークン $\mathbf{s}_1$ には音響イベントの骨格情報が高い密度で圧縮されていると考えられ，構造的な異常の変化を捉えやすい可能性がある．

次章以降では，この WavTokenizer によって得られた単一のトークン列 $\mathbf{S}$ および再構成波形 $\hat{\mathbf{x}}$ を用いた具体的な異常検知手法について述べる．

3. 離散トークンに基づく異常検知フレームワーク

本章では，事前学習済みの WavTokenizer を特徴抽出器として用いた，教師なし異常検知のフレームワークについて述べる．提案手法の全体概要を図 2 に示す．本研究では，正常データのみを用いて学習を行う設定（One-class classification）を前提とし，異常度の算出手法として，信号レベルの復元誤差に着目した手法（上段）と，トークン列の統計的規則性に着目した手法（下段）の 2 種類を提案・比較する．

3.1 共通処理：音響信号のトークン化

まず，入力された音響信号 $\mathbf{x}$ に対する前処理として，WavTokenizer を用いたトークン化を行う．一般に RVQ ベースの手法は複数のトークン列 $\mathbf{s}_1, \ldots, \mathbf{s}_{N_q}$ を出力するが，本研究では WavTokenizer の代表的な構成である単一層量子化（$N_q = 1$）の事前学習済みモデルを採用する．これにより，入力波形は追加の選別操作を行うことなく，音響情報の骨格を集約した単一の量子化インデックス列 $\mathbf{s}_1$ へと変換される．

本研究では，この第 1 層の出力をそのまま解析対象のトークン列 $\mathbf{S}$ として定義する（すなわち $\mathbf{S} = \mathbf{s}_1$）．具体的には，$\mathbf{S}$ を時間方向に展開して $\mathbf{S} = (s_1, s_2, \ldots, s_L)$ と表記する．ここで L は系列長，s_t は時刻 t におけるトークン ID（$s_t \in \{1, \ldots, V\}$，V は語彙数）である．

この処理により，高次元な時系列信号 $\mathbf{x} \in \mathbb{R}^T$ は，より低次元な離散整数系列 $\mathbf{S} \in \{1, \ldots, V\}^L$ $(L \ll T)$ へと変換される．この $\mathbf{S}$ が，以下の 2 つの手法における共通の入力となる．

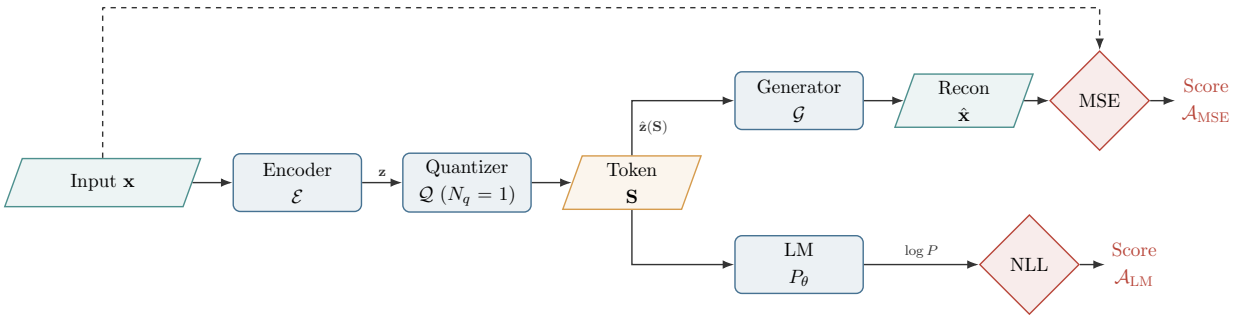

Fig. 2　提案手法の概要．WavTokenizer $(\mathcal{E}, \mathcal{Q})$ により抽出された単一のトークン列 $\mathbf{S}$ に対し，Generator $\mathcal{G}$ による再構成誤差（MSE）と言語モデル P_θ による対数尤度（NLL）の 2 つのアプローチで異常スコアを算出する．

3.2　手法 1：再構成誤差に基づく検知 (Recon-MSE)

1 つ目のアプローチは，WavTokenizer をオートエンコーダとして見立て，その再構成誤差（Reconstruction Error）を異常スコアとする手法である．WavTokenizer は大規模な一般的音声データセットで学習されているが，特定の機械音が発する定常的なノイズや振動音に対しては，適応が不十分な場合がある．しかし，正常な機械音は比較的単純な構造を持つため，汎用モデルであってもある程度の再構成が可能であると仮定する．

一方で，突発的な異音や未知の異常パターンが含まれる場合，その成分は事前に学習されたコードブックで表現できず，デコーダによる再構成波形 $\hat{\mathbf{x}}$ と元の波形 $\mathbf{x}$ との間に乖離が生じると考えられる．そこで，入力波形 $\mathbf{x}$ と，トークン列 $\mathbf{S}$ から再構成された波形 $\hat{\mathbf{x}}$ との間の平均二乗誤差（MSE）を異常スコア $\mathcal{A}_{\mathrm{MSE}}$ として定義する．ここで，再構成波形 $\hat{\mathbf{x}}$ は，トークン列 $\mathbf{S}$ に対応するコードブックベクトル（量子化潜在表現）を $\hat{\mathbf{z}}(\mathbf{S})$ としたとき，デコーダ $\mathcal{G}$ を用いて以下のように表される．

$$\hat{\mathbf{x}} = \mathcal{G}(\hat{\mathbf{z}}(\mathbf{S})) \tag{9}$$

したがって，異常スコアは次式で計算される．

$$\mathcal{A}_{\mathrm{MSE}}(\mathbf{x}) = \frac{1}{T}\sum_{i=1}^{T}(x_i - \hat{x}_i)^2 \tag{10}$$

この手法は，追加の学習を一切必要とせず，事前学習済みモデルの推論のみで異常検知が可能であるという利点を持つ．

3.3　手法 2：言語モデルに基づく検知 (Token-LM)

2 つ目のアプローチは，トークン列 $\mathbf{S}$ を自然言語の文章とみなして言語モデル（Language Model: LM）を学習させ，その生成確率（尤度）を異常スコアとする手法である．正常な機械の動作音は，回転や往復運動などに起因する一定の時間的規則性（文法）を持っているはずである．この規則性を LM に学習させることで，文法的に不自然なトークン列（異常音）を検知する．このアプローチは，自然言語処理における GPT[8] などの自己回帰モデルや，Transformer[9] を用いた時系列異常検知と同様の枠組みである．

具体的には，学習用データセットに含まれる正常データ（Source ドメインおよび少数の Target ドメイン）から抽出されたトークン列 $\mathbf{S}$ を学習データとし，Transformer ベースの自己回帰モデル（Causal Transformer）を学習させる．モデルは，時刻 t までのトークン列 $s_{<t}$ を条件として，次のトークン s_t の条件付き確率 $P(s_t|s_{<t})$ を予測するように訓練される．異常検知時には，テストデータのトークン列に対する負の対数尤度（Negative Log-Likelihood: NLL）を計算し，

これを異常スコア $\mathcal{A}_{\mathrm{LM}}$ とする．

$$\mathcal{A}_{\mathrm{LM}}(\mathbf{S}) = -\frac{1}{L}\sum_{t=1}^{L}\log P(s_t|s_{<t};\theta) \qquad (11)$$

ここで θ は正常データのみで学習された LM のパラメータである．本手法では，限られた正常データでの過学習を防ぐため，パラメータ共有（Weight Tying）や層数の削減を行った軽量な Transformer モデルを採用する．これにより，個別の音響特性ではなく，トークン列の普遍的な遷移パターンのみを学習させることを意図している．

4. 実験評価

本章では，提案した 2 つの手法（Recon-MSE, Token-LM）を実際の機械音データセットに適用し，DCASE 公式ベースラインとの比較を通してその有効性を評価する．

4.1 実験設定

4.1.1 データセット

実験には DCASE 2022 Task 2 ”Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques” の開発データセットを用いた．評価対象は 7 種類の機械（ToyCar, ToyTrain, Fan, Gearbox, Bearing, Slider, Valve）であり，各機械について 3 つのセクション（section 00–02）のデータを用いた．

本タスクの特徴として，学習データと同一の動作条件で収録された「Source ドメイン」と，速度や雑音条件などが変化した「Target ドメイン」が存在する．本実験では，Token-LM の学習には Source ドメイン（および少数の Target ドメイン）の正常データを用い，Recon-MSE は事前学習済みモデルのみ（Zero-shot）で推論を行った．

4.1.2 事前学習モデル

WavTokenizer の公式事前学習済みモデルの中から，学習データの規模やモデル容量の異なる以下の 2 つのモデルを選定し，その性能差を検証した．

- `small_600`: LibriTTS 等の標準的な音声データセットで学習された軽量モデル．
- `large_unify_40`: 音声・音楽・環境音を含む大規模な統合データセットで学習された汎用モデル．

なお，いずれのモデルも量子化層数は単一（$N_q = 1$），語彙サイズは 4096，サンプリングレートは 24kHz の設定である．

4.1.3 比較手法

本研究では，以下の (1), (2) の 2 つのアプローチを提案し，公式ベースラインと比較する．

(1) Recon-MSE (Zero-shot)
WavTokenizer をオートエンコーダとして用い，入力波形と再構成波形の平均二乗誤差（MSE）を異常スコアとする手法．本手法は事前学習済みモデルの推論能力のみに依存し，DCASE データセットを用いた追加学習（Fine-tuning）は一切行わない．

(2) Token-LM (Train from scratch)
第 1 層トークン列の遷移確率を学習した言語モデルを用い，テストデータの負の対数尤度（NLL）を異常スコアとする手法．モデルには軽量な Transformer Decoder（層数 3，埋め込み次元 256，ヘッド数 4，ブロック長 256）を採用した．学習データとして Source ドメイン（および少数の Target ドメイン）の正常データのみを用い，30 エポックの学習を行った．

Baseline-AE
比較対象として，DCASE 2022 Task 2 公

Table 1　Source ドメインにおける AUC

Machine	Base	Recon-MSE		Token-LM	
	AE	*Small*	*Large*	*Small*	*Large*
ToyCar	**0.917**	0.545	0.566	0.552	0.502
ToyTrain	**0.770**	0.558	0.580	0.446	0.470
Bearing	**0.570**	0.515	0.533	0.541	0.431
Fan	**0.790**	0.627	0.671	0.558	0.593
Gearbox	**0.692**	0.588	0.543	0.545	0.531
Slider	**0.788**	0.526	0.583	0.627	0.628
Valve	**0.521**	0.304	0.292	0.478	0.523
Average	**0.721**	0.524	0.538	0.535	0.525

式の標準的なオートエンコーダの結果も併記する [7]．これは対数メルスペクトログラムを入力とし，正常音の分布を学習するモデルである．なお，ToyCar および ToyTrain のベースライン値については，該当データセットのベンチマーク結果 [10] を参照した．

4.1.4　評価指標

主たる指標として，Source/Target それぞれのドメインにおける AUC (Area Under the Curve) を用いる．また，誤検知を低く抑えることが求められる実運用を想定し，部分 AUC（pAUC, FPR$\leq p$）についても評価する．ここで $p = 0.1$ とし，次式で算出する．

$$\mathrm{pAUC} = \frac{1}{p}\int_0^p \mathrm{TPR}(x)\,dx \tag{12}$$

本定義において，ランダム推論時の期待値は 0.5 となる．したがって，スコアが 0.5 を上回っていれば，低誤検知率領域においてランダム以上の有意な検知能力を有していると判断できる．

4.2　実験結果および考察

各機械種別における異常検知性能を表 1（Source ドメイン），表 2（Target ドメイン），表 3（pAUC）に示す．以下，これらの結果に基づき，ドメインシフトへの頑健性および手法ごとの特性について考察する．

Table 2　Target ドメインにおける AUC．太字は Baseline を上回ったスコアを示す．

Machine	Base	Recon-MSE		Token-LM	
	AE	*Small*	*Large*	*Small*	*Large*
ToyCar	0.366	**0.481**	**0.481**	**0.507**	**0.479**
ToyTrain	0.264	**0.507**	**0.513**	**0.511**	**0.532**
Bearing	**0.590**	0.517	0.509	0.518	0.537
Fan	0.492	**0.654**	**0.606**	**0.575**	**0.573**
Gearbox	**0.628**	0.562	0.601	0.558	0.485
Slider	0.490	0.470	**0.496**	**0.567**	**0.550**
Valve	**0.499**	0.372	0.304	0.485	0.443
Average	0.476	**0.509**	**0.501**	**0.532**	**0.514**

Table 3　平均 pAUC（$p = 0.1$）

Machine	Base	Recon-MSE		Token-LM	
	AE	*Small*	*Large*	*Small*	*Large*
ToyCar	0.528	0.099	0.118	**0.127**	0.053
ToyTrain	0.506	**0.081**	0.070	0.075	0.078
Bearing	0.522	**0.201**	0.190	0.073	0.080
Fan	0.580	0.167	**0.229**	0.131	0.173
Gearbox	0.587	0.160	**0.239**	0.111	0.097
Slider	0.560	0.183	0.193	**0.206**	0.201
Valve	0.504	0.017	0.027	**0.069**	0.058
Average	0.541	0.130	**0.152**	0.113	0.106

まず，学習データと同一条件である Source ドメインの結果（表 1）に着目すると，Baseline-AE が全体的に高い性能を示した．これは，タスク専用に学習されたモデルがドメイン内の特徴分布を精緻に捉えているのに対し，汎用モデルを用いた Recon-MSE（Zero-shot）や限定的な学習を行った Token-LM では，正常音の分布への適応が十分ではなかったためと考えられる．

一方で，ドメインシフトが発生する Target ドメイン（表 2）では異なる傾向が見られた．Baseline-AE の性能は低下し，特に ToyTrain や ToyCar では検知が困難な状況となっている．これに対し，提案手法はいくつかの条件で Baseline と比較して良好な結果を示した．具体的には，Fan において Recon-MSE（Small）が Baseline を上回ったほか，ToyCar や ToyTrain においても Token-LM 手法が 0.5 以上のスコアを維持した．これらの結果は，WavTokenizer の離散表現が，音響条件の変動に対して一定の汎化性能を有している可能性を示唆している．

表 3 に示す通り，Baseline-AE は pAUC にお

いてもランダム基準を上回り，安定した性能を示している．対して，提案手法の平均値はランダム水準を下回る結果となった．これは，汎用モデルを用いた Zero-shot 推論（Recon-MSE）や限定的な学習（Token-LM）では，正常音分布の裾野を十分に学習できず，一部の正常データを高く異常判定してしまう（誤検知する）傾向があることを示している．実用化にはファインチューニング等の対策が不可欠である．しかしながら，手法間の相対的な特性差に着目すると，重要な知見が得られる．Fan や Bearing などの定常音では Recon-MSE が Token-LM を上回る一方，Valve のような非定常音では，Recon-MSE に対して Token-LM が相対的に高い値を維持した．絶対値としては低いものの，これは波形 MSE が苦手とする非定常音を，LM の文脈情報が補完するという相補的な関係を示唆しており，今後の精度向上に向けた重要な指針となる．

5. 結言

本研究では，音声生成分野で注目される離散音声トークナイザ（WavTokenizer）を異常検知に適用し，その適用可能性について基礎的な検討を行った．DCASE 2022 Task 2 を用いた評価実験の結果，汎用的な事前学習済みモデルであっても，Fan 等の定常音に対しては再構成誤差法（Recon-MSE）が有効に機能し，ドメインシフト環境下においてもベースラインと同等以上の性能を示すケースが確認された．一方で，Valve 等の非定常音に対しては波形 MSE による検知が困難であったが，トークン列の言語モデル化（Token-LM）がその検知漏れを補完する傾向が見られ，波形レベルでは捉えきれない異常を文脈として検知できる可能性が示唆された．

今後は，残差量子化（RVQ）の第 2 層以降を活用した詳細な特徴抽出や，機械音データによるファインチューニング，および位相影響を受けにくい評価指標の導入などが，精度向上のために重要になると考えられる．

謝辞

本研究の一部は、JST【ムーンショット型研究開発事業】【JPMJMS2292】の支援を受けたものである。

参考文献

1) A. van den Oord, O. Vinyals, and K. Kavukcuoglu, “Neural Discrete Representation Learning,” in *Proc. NIPS*, 2017.
2) N. Zeghidour et al., “SoundStream: An End-to-End Neural Audio Codec,” *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, vol. 30, pp. 495–507, 2021.
3) A. Défossez et al., “High Fidelity Neural Audio Compression,” *arXiv preprint arXiv:2210.13438*, 2022.
4) Z. Borsos et al., “AudioLM: a Language Modeling Approach to Audio Generation,” *IEEE/ACM Trans. Audio, Speech, and Lang. Proc.*, vol. 31, pp. 2523–2533, 2023.
5) S. Ji et al., “WavTokenizer: an Efficient Acoustic Discrete Codec Tokenizer for Audio Language Modeling,” *arXiv preprint arXiv:2408.16532*, 2024.
6) R. Chalapathy and S. Chawla, “Deep Learning for Anomaly Detection: A Survey,” *arXiv preprint arXiv:1901.03407*, 2019.
7) K. Dohi et al., “Description and Discussion on DCASE 2022 Challenge Task 2: Unsupervised Anomalous Sound Detection for Machine Condition Monitoring Applying Domain Generalization Techniques,” in *Proc. DCASE Workshop*, 2022.
8) A. Radford et al., “Language Models are Unsupervised Multitask Learners,” *OpenAI Blog*, vol. 1, no. 8, 2019.
9) A. Vaswani et al., “Attention Is All You Need,” in *Proc. NIPS*, 2017.
10) N. Harada et al., “ToyADMOS2: Another Dataset of Miniature-Machine Operating Sounds for Anomalous Sound Detection under Domain Shift Conditions,” in *Proc. DCASE Workshop*, pp. 1–5, 2021.