一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

既存データセットとの意図しない重複を避ける 環境音評価データセットの半自動構築法

岸　　秀[†]　高道慎之介[†,††]　滝沢　　力[†††]　金森　勇介[††]　砺波　紀之[††††]
永瀬亮太郎[†††††]井本　桂右[†††††]岡本　悠希[††]

† 慶應義塾大学
†† 東京大学
††† 京都産業大学
†††† 日本電気株式会社
††††† 立命館大学
†††††† 京都大学
E-mail: †{minorukishi1091,shinnosuke __ takamichi}@keio.jp

あらまし　本稿では，環境音評価データセットの半自動構築手法を提案する．機械学習モデルの性能を正確に評価するために，学習データと重複しない評価データセットの使用が不可欠である．しかし，環境音分野においては，Web から収集した音データや，既存コーパスを継承した音データを使用することが多いため，利用者が意図せずとも，学習データとの重複が発生してしまう．さらに，新たにデータを収集するには多大な時間的・金銭的コストが必要となる．そこで本研究では，重複しない環境音評価セットを，効率的かつ半自動的に構築する方法を提案する．

Minoru KISHI[†], Shinnosuke TAKAMICHI[†,††], Riki TAKIZAWA[†††], Yusuke KANAMORI[††], Noriyuki TONAMI[††††], Ryotaro NAGASE[†††††], Keisuke IMOTO[††††††], and Yuki OKAMOTO[††]

† Keio University
†† The University of Tokyo
††† Kyoto Sangyo University
†††† NEC corporation
††††† Ritsumeikan University
†††††† Kyoto University
E-mail: †{minorukishi1091,shinnosuke __ takamichi}@keio.jp

1. はじめに

深層学習技術の進展に伴い，環境音の合成や認識といったタスクの性能が著しく向上している [1]〜[5]．この技術の発展には，開発されたモデルを客観的かつ信頼性の高い評価データセットを用いて適切に評価することが不可欠である．しかし，環境音評価の分野において現在利用されている多くの環境音データセットは，Web から収集された音データや，過去のデータセットを継承・加工したデータが多くを占めている [6]〜[9]．この結果，利用者が意図せずとも，モデルの学習に用いたデータと評価に用いるデータとの間で重複が生じてしまう可能性がある．このような重複はモデル評価の公正さを低下させてしまう．このような学習データと評価データの意図しない重複は，画像認識などの分野でもモデルの性能を過大評価させる要因として指摘されており，厳密な管理が求められる [10]．これを回避する最も確実な方法は，独自に音を収録することである．しかし，環境音は発生源や録音環境が極めて多様であり，人手で収録を行い十分に多様かつ大規模なサンプルを収集することは，時間的・金銭的コストの観点から非現実的である．

本稿では，この課題に対し，高い信頼性と低コストを両立した評価環境の実現を目指す．具体的には，以下の 3 つのアプローチを統合した，環境音評価データセットの半自動構築手法を提案する．まず，意図しない重複を避けるため，公開期間に基づく厳格なフィルタリングを導入する．Web 上の大規模モデルは学習データの詳細が不明な場合も多いが，本手法では評価対象モデルの公開日やデータ収集のカットオフ日よりも後に

公開された音データを含む動画のみを収集対象とする．次に，構築コストの削減のため，大規模言語モデル（LLM）と環境音認識モデルを組み合わせた半自動パイプラインを構築する．半自動パイプラインは図 1 に示す．音響イベントラベルである初期シード語（例: “laughter”）を入力とし，LLM を用いて関連語を自動生成する．さらに収集プロセスの負例から除外キーワードを動的に抽出して検索クエリに反映させる適応的クエリ更新戦略を提案する．これにより，人手による膨大な試行錯誤を介さずに，所望する環境音を効率的に収集することが可能となる．最後に，データの多様性と品質の確保のため，音響イベント検出モデルによる自動スクリーニングと，信頼度に基づく正例・負例の自動判定を行う．これにより，Web データ特有の偏り（ロングテール分布）を是正し，評価用として適切な品質を持つデータセットを低コストで構築する手法を確立する．なお，提案手法を基に構築したデータセットは，環境音生成モデルのテキスト-環境音関連度を予測する国際チャレンジ XACLE Challenge 2026 [11] の test set として採用されており，公式ページ(注1)から入手可能である．

2. 関 連 研 究

2.1 既存データセットの音データ収集方法

環境音認識や合成の研究には大規模データセットが不可欠であり，代表例として YouTube から収集された AudioSet [12] や AudioCaps [6] が挙げられる．AudioSet は 632 クラスのオントロジーに基づく約 200 万の動画クリップに対し，候補選別と人手判定を経てラベルが付与された．AudioCaps はこのラベル付けされたクリップの一部のサンプルにキャプションを付与したものである．

加えて，分類タスクでは ESC-50 [13] や UrbanSound8K [14]，キャプション生成では Clotho [15] や WavCaps [16] が標準的である．また，映像・音声の対応関係を重視した VGGSound [17] や，音の共有プラットフォームである Freesound(注2)を音源とする FSD50K [18] なども提案されている．これらは Web データを主な音源とする点で本研究と共通する．しかし，公開から時間が経過しており，近年の大規模モデルの学習データとして意図せず取り込まれている可能性がある．

2.2 既存データセットの課題

2.2.1 意図しないデータリーク

評価の公平性を損なう要因として，主に以下の 3 つの意図しないデータリークが挙げられる．

Web クロールによるリーク. CLAP [19] や Qwen-Audio [**?**] 等のモデルは大規模 Web データを学習に用いるため，カットオフ日以前に公開されたベンチマークが意図せず混入する可能性がある．特に，動画共有サイトでは同一コンテンツが別 ID で再アップロードされる事例があり，ID に基づく分割では学習・テストデータ間の重複を完全に防げないリスクがある．

データセット間の継承関係によるリーク. 既存の大規模データセット（親）の一部を流用したデータセット（子）をテストに用いる際，モデルが親データセットで事前学習されているとリークが生じる．例えば，AudioCaps [6] のテストデータは AudioSet の学習データから抽出されている．そのため，AudioSet で事前学習した AST [20] を AudioCaps で評価すると，既知のデータに対する推論となり，汎化性能の正当な評価が困難となる．

学習データの不透明性によるリーク. 学習データの詳細が非公開であるモデルは，ベンチマークの混入検証が困難である．例えば，Whisper [21] 等のモデルは広範な Web データを学習しているとされるが，その詳細は不明である．仮に評価データが学習過程に含まれていた場合，それはデータリークであり，やはり汎化性能の正確な測定を妨げる要因となる．

2.3 莫大なコスト

多くの音響データセット構築の過程で人手の介入が必要となる．例えば収集した音響データを聴取し，所望のイベントが含まれているかを検証する．このような人的資源を要する作業はデータセット構築を妨げる大きな要因である．

2.4 音響イベントラベルのサンプル数の偏り

AudioSet [12] をはじめとする主要なデータセットでは，話し声や音楽といった一般的なクラスのサンプル数は豊富である一方，ガラスの割れる音や特定の動物の鳴き声などの希少クラスは極端に少ない．評価において，音響イベントのサンプル数の偏りを無視すると見かけ上の性能が高くなり，希少な音響イベントに対する性能が低いモデルの開発を招く恐れがある．

3. 提 案 手 法

3.1 要請される条件

2.2.1 節で議論した既存データセット構築の問題を回避するためには，データセット構築過程に条件を設ける必要がある．本研究では Web からのクローリングを前提とし，以下の 3 つの条件を満たすアルゴリズムによってデータ収集を行う．

公開期間による厳格なフィルタリング． 学習データの不透明性への対策として，モデルの公開日または学習カットオフ日より後に公開された動画のみを収集対象とする．これにより，再アップロード等の例外を除き，モデルがそのデータを未学習であることを保証する．

イベントラベルごとのサンプル数の均一化． Web 上の音データに含まれる音響イベントの音楽や話し声への偏り [12] を是正するため，単純なランダムサンプリングではなく，希少な環境音も含めた全クラスで均一なデータ量を確保する．

データ収集の半自動化． 多様な音響イベントの網羅と継続的なデータ取得を実現するため，時間的・金銭的コストを抑制する．膨大な Web データから目的の音源を効率的に抽出し，最小限の人手で品質を担保することで，人手による高コストな収録・確認作業を代替する．

3.2 収集アルゴリズムのフロー

本節では，Web 上の動画共有プラットフォーム（本研究では YouTube を対象とする）から，目的とする環境音データを効率的かつ高精度に収集するための半自動構築アルゴリズムについて述べる． 図 1 に示すように本手法は，初期シード語を入力と

(注1)：https://xacle.org/
(注2)：https://freesound.org/

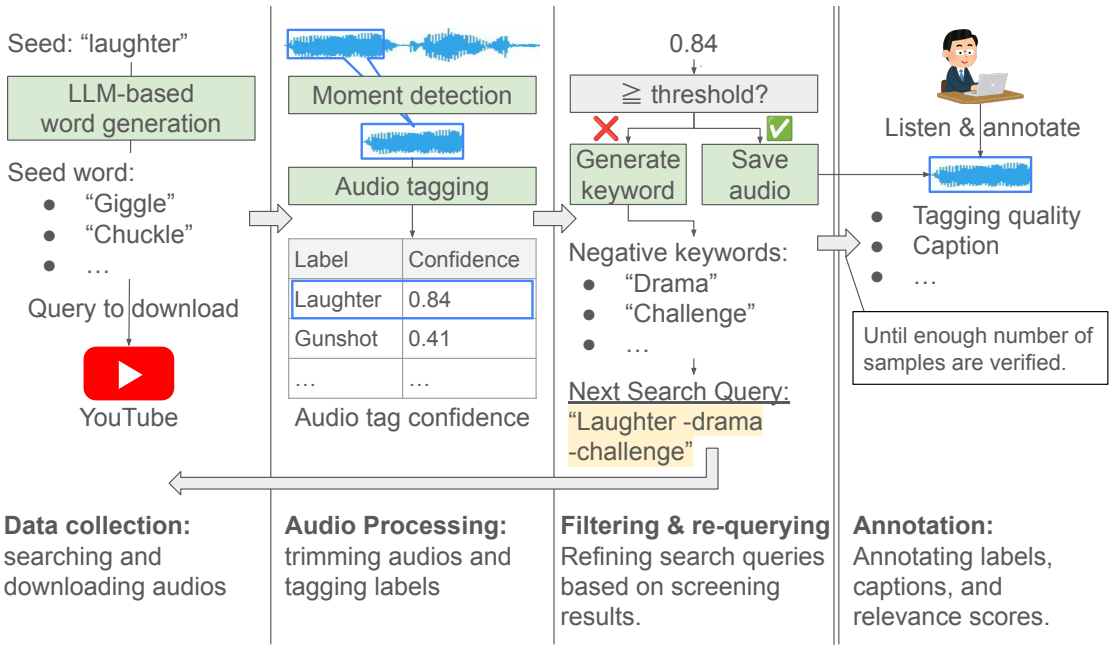

図 1　データセット構築フロー

し，検索，ダウンロード，音響イベント検出モデルによるスクリーニング，および検索クエリの適応的な更新を反復することで，データセットを構築する．

3.2.1　アルゴリズム概要

提案する収集パイプラインの全体像は，以下の 4 つの主要なステップから構成される．

（1） **初期化と関連語の生成**: 入力されたシード語（例: “laughter”）に基づき，大規模言語モデル（LLM）を用いて，検索の候補となり得る関連語リスト（related words）を生成する．シード語と関連語ともに動画検索クエリとして使用するが，この関連語はシード語のみでは検索結果が枯渇した場合の代替クエリとして機能する．

（2） **メタデータ検索とフィルタリング**: 現在の検索クエリを用いて動画検索を行い，候補となる動画リストを取得する．このとき，検索結果を公開日の新しい順（降順）でソートして取得する．これにより，3.1 節で述べた公開期間による厳格なフィルタリングの条件を満たす動画を優先的に効率よく収集できる．取得したリストからは，既定の期間より前に公開された動画の ID 及び既に収集済みの ID を除外することで，データの重複を防止する．

（3） **音データ取得と自動スクリーニング**: 選定された動画から音データを規定フォーマットでダウンロードし，固定長に切り出す．続いて，事前学習済みの音響イベント検出モデルを用いて各クリップに音響イベントラベルおよび信頼度を付与する．信頼度が閾値 τ（例: 0.3）を超えるクリップのみを正例（positive）としてデータセットに保存し，それ以外を負例（negative）として破棄する．

（4） **適応的クエリ更新**: スクリーニングの結果，所望するクリップ数を集められなかった場合，検索クエリを自動更新して (1)-(3) を再度実行する．詳細は次項で述べる．

3.3　適応的クエリ更新戦略

提案アルゴリズムにおける動画検索の課題は，音響イベント名を検索クエリにしたとき，(1) 当該イベントに無関係の動画がヒットすることと (2) 動画が一切ヒットしないことである．そこで，収集したデータの自動スクリーニング結果に応じて，除外キーワードによる絞り込みと関連語による探索範囲の拡大を切り替える戦略を採用する．

3.3.1　負例からの除外キーワード抽出

スクリーニングにおいて信頼度が閾値を下回った負例の動画タイトルには，目的音とは異なる音源（例: “Laughter” 検索時の “Try not to laugh challenge” のような BGM 動画など）が含まれる事例がある． 本手法では，負例と判定された動画群のタイトルを解析し，頻出する単語を除外キーワード（negative keywords）として抽出する．本実装では，負例タイトルから単語の出現頻度をもとに決定する．ストップワードや極端に少ない頻度の語以外の抽出された候補語を文書頻度（DF）順にソートし，上位語を除外キーワードとして利用する．次回の検索クエリには，現在の検索語に加え，これらの除外キーワードをマイナス検索（not 条件）として付与する（例: “Laughter -challenge -compilation”）．これにより，イテレーションを重ねるごとに検索精度を向上させる．

3.3.2　関連語への切り替え

同一のシード語に対して除外キーワードを追加し続けても新規の動画が見つからなくなった場合（すなわち，動画ヒット数が 0 件），あるいは一定回数動画検索から自動スクリーニングまで行っても目的数のデータが得られなかった場合は，探索範囲が限界に達したと判断する． この際，初期化ステップで生成した関連語リストから次の単語（例: “Giggle” や “Chuckle”）を取り出し，検索クエリの語を関連単語に切り替える．これにより，単一の単語では網羅できない多様な音データを収集可能とする．

4. 実験的評価

4.1 実験条件

提案アルゴリズムを用いて環境音評価データセットを構築した結果を報告する．なお，収集したデータセットを XACLE Challenge 2026 の test set として利用するためのキャプションおよび主観評価値の付与処理については，Appendix1. を参照されたい．

4.1.1 収集アルゴリズムにおける使用モジュール

提案手法の音データ収集に際し，以下の 3 つの学習済みモデルを用意した．

- **AM-DETR [22]**(注3)：Transformer [23] エンコーダ・デコーダ構造に基づくオーディオ・グラウンディングモデルである．本実験では `lighthouse-emnlp2024/AM-DETR` のデフォルトの重みを使用した．本手法においては，入力されたシード語（テキスト）に対応する音響イベントが，収集した動画のどの時間区間に存在するかを推定し，当該区間を切り出すために使用する．
- **EAT [24]**(注4)：画像認識モデル ViT [25] の構造を環境音認識に応用した Transformer ベースのモデルである．AudioSet [12] の約 200 万データで学習された `EAT-base_epoch30_finetune_AS2M` の重みを使用した．切り出された音データに対して音響イベントタグを付与し，目的とするイベントが含まれているかを判定する最終的な自動スクリーニング（正例・負例の選別）に使用する．
- **GPT-4o-mini**：OpenAI によって開発された大規模言語モデル (LLM) である．本手法では，初期シード語からの関連語リストの生成，および収集プロセスにおいて負例から抽出された除外キーワードを反映した検索クエリの適応的な更新に使用する．

4.1.2 収集反復および終了条件

本実験におけるデータ収集プロセスでは，効率性と計算コストのバランスを考慮し，以下の反復ルールおよび終了条件を設定した．同一のシード語に対して除外キーワードを追加し続けても新規の動画が見つからない場合，あるいは動画検索から自動スクリーニングまでの工程を 3 回繰り返しても目的数のデータが得られなかった場合は，現在の検索語での探索範囲が限界に達したと判断し，生成された関連語へと検索クエリを切り替える設定とした．さらに，各イベントラベルにおける収集プロセスは，自動スクリーニングを通過した正例データが 40 サンプル以上集まった時点で収集完了とした．また，無限ループや過度な API 消費を防止するため，一つのイベントラベルに対する総イテレーション回数の上限を最大 12 回に設定した．この上限に達した場合は，確保できたサンプル数が規定の 40 個未満であっても，当該イベントに対する収集を打ち切ることとした．

4.1.3 収集データ対象期間

データ収集の対象期間は，2025 年 6 月以降に公開された動画とした．これは，評価対象となる既存の大規模モデル（例えば 2025 年以前に学習が完了しているモデル）の学習データと収集データの意図しない重複を排除するためである．

4.1.4 データソース

音響データの収集元として，世界最大級の動画共有プラットフォームである YouTube を選定した．YouTube は，日常的な環境音から希少な音響イベントまで多岐にわたる動画データを含んでおり，実環境における多様な音響シーンを網羅的に収集するために適している．選定された動画からのデータ取得においては，規定フォーマットとしてサンプリング周波数 32 kHz，RIFF WAV ファイル形式を採用した．最終的に 4.1.1 節の AM-DETR の推論結果をもとにそのイベントが最大時間含まれるような 10 秒間を切り出した．

4.1.5 対象イベントラベル

収集対象とする音響イベントのカテゴリ選定には，環境音キャプション生成タスクの標準的なデータセットである AudioCaps [6] を採用した．具体的には，AudioCaps に含まれる 68 種類のイベントクラスを対象とし，これらを初期シード語として本手法による収集プロセスを実行した．これにより，既存データセットとのドメイン分布の比較や，同一カテゴリにおける品質の比較検証を可能とした．また，イベントラベルの分布の偏りが大きくなるのを防ぐため各イベント最低 5 サンプル含むようにデータを収集した．

4.1.6 自動スクリーニングの精度の算出方法

収集された音データのうち，ターゲットとする音響イベントが正しく含まれている割合を適合率（Precision）として定義する．自動スクリーニングを通過したデータ集合を $\mathcal{S}$，その中で人手による聴取確認によって正例と判定された集合を $\mathcal{S}_{\text{pos}}$ とし，精度は $|\mathcal{S}_{\text{pos}}|/|\mathcal{S}|$ により算出する．なお，ここでの正例判定（$\mathcal{S}_{\text{pos}}$ の決定）においては，作業者に対し「対象の音響イベントが時間区間の半分以上を占めており，かつ背景に話し声や音楽などの他のイベントが含まれていない場合のみを正例とする」という厳格なインストラクションに基づき確認を行った．

4.1.7 データセットの分布の比較

提案手法により半自動構築されたデータセットの品質およびドメインの妥当性を検証するために，既存の標準的なデータセットである AudioCaps との分布比較を行う．

分布の可視化． 音データの特徴抽出器には AST の 13 層目のフレームごとの出力された埋め込み表現を平均することで各サンプルに対応する 1 つの埋め込み表現を得た．AST の重みは `Full AudioSet, 10 tstride, 10 fstride, with Weight Averaging (0.459 mAP)`(注5)を使用した．また，可視化を行う際は t-SNE を用いて 2 次元にプロットを行った．

分布間距離． 提案したデータセットと既存のデータセット AudioCaps との間の音響特徴量分布の類似性を測るために，以下の 3 つの距離尺度を使用する．

（1） **Fréchet Distance (FD) [26]**：特徴空間における 2 つのガウス分布間の距離を測定する．本実験では分布の可視化でも

（注3）：https://github.com/lighthouse-emnlp2024/AM-DETR

（注4）：https://github.com/cancellieri/EAT

（注5）：https://www.dropbox.com/s/ca0b1v2nlxzyeb4/audioset_10_10_0.4593.pth?dl=1

使用した AST による埋め込み表現を用いて計算した．

（2） **Fréchet Audio Distance (FAD) [27]**：FD と同様の概念だが，VGGish などの音響モデルから得られる埋め込み特徴量を用いて計算され，音質の劣化やドメインの不一致に敏感である．

（3） **KL ダイバージェンス (Kullback-Leibler Divergence)**：2 つの確率分布間の差異を情報理論的に測定する指標である．本実験では VGGish モデルから得られる特徴量分布を用い，提案したデータセット（Tgt）と既存データセット AudioCaps（Ref）間での分布の乖離を定量的に評価するために用いる．ここで，KL(Ref ‖ Tgt) は既存データセットの分布に対する提案データセットの網羅性を表し，KL(Tgt ‖ Ref) は既存データセットからの逸脱度を表す．

4.1.8 モデルの比較評価

提案手法により構築されたデータセットが，実際のモデル学習や評価において有用であるかを検証するために，RELATE [28] を使用する．本実験では XACLE Challenge 2026 のベースラインである事前学習済みの重み(注6)を使用した．RELATE は，音データとテキストの関連性を評価するモデルであり，本実験ではこのモデルを用いて既存データセットと提案データセットのそれぞれに対するスコア算出した．これらのスコアと主観評価値との相関および誤差を計算しデータセットごとの比較分析を行った．データセットは XACLE Challenge 2026 dataset validation set の自然音，XACLE Challenge 2026 dataset test set の自然音および XACLE Challenge 2026 dataset test set 合成音を使用した．XACLE Challenge 2026 dataset test set 合成音は 2,000 サンプルのうちランダムで 1,000 サンプルを選択した．また，XACLE Challenge 2026 dataset validation set の自然音は AudioCaps の test set から抽出したものである．

相関係数 (SRCC, LCC, KTAU)：モデルの予測スコアと正解スコアの間の相関の強さを測定する指標である．スピアマン順位相関係数 (SRCC) およびケンドールの順位相関係数 (KTAU) は順位関係の単調性を，ピアソンの積率相関係数 (LCC) は線形的な相関関係を評価する．いずれの指標も，値が 1 に近いほどモデルの予測が人間の感覚や正解データと一致していることを示す．

平均二乗誤差 (MSE)：予測スコアと正解スコアとの差の二乗平均であり，モデルの予測精度の高さを評価する．値が 0 に近いほど予測誤差が小さいことを示す．

4.2 実 験 結 果

4.2.1 収集データの歩留まりと精度

提案手法を用いて自動収集およびスクリーニングを実施した結果，その有効性とラベル付けの精度を定量的に評価した．自動スクリーニングによって正例と判定されたサンプル数は 1457 件であった．これらに対し，人手による聴取確認を行った結果，931 件が最終的に正例として選定された．音響イベントごとのサンプル数は Appendix2. を参照されたい．この結果より，自動スクリーニングを通過したデータのうち約 63.9% が，実際に目的とする音響イベントを含み，かつ音楽や話し声などの背景音がないことが確認された．完全な手動収録と比較して，提案手法は人手による作業を最終確認のみに限定できるため，大幅なコスト削減と効率化が実現できたと結論付けられる．

4.2.2 音データ収集に要したコスト

本提案手法によるデータ収集において発生した時間的および金銭的コストについて，以下の各項目ごとに算出・分析を行った．

コードの実行時間. データ収集および自動処理プログラムの実行には，総計で約 60 時間を要した．このときダウンロードおよび処理したサンプル数は 38887 サンプルである．本処理は，Intel Core i7-14700KF（20 コア 28 スレッド）を搭載した計算機上で実行した．なお，動画ダウンロード工程のみ CPU 並列数 8 に設定し，その他の処理は逐次的に実行した．これは計算機による自動処理であり，人的リソースを占有しない時間である．

人手によるスクリーニング工数. 自動スクリーニングを通過した 1457 サンプルに対する人による聴取確認を行った．作業効率は約 100 サンプルあたり 1 時間であり，総作業時間は 15 時間であった．これは，数千規模のデータをゼロから収集・選定する場合と比較して極めて少ない工数である．

API 使用料. 大規模言語モデル GPT-4o-mini を，検索クエリの拡張に用いる関連語リストの生成のみに使用した．具体的には，対象となる全 68 種類のイベントに対し，初期シード語を入力として各 1 回ずつ推論を行った．これに伴うトークン消費量は，入力が約 1.2 万トークン，出力が約 0.4 万トークン（合計約 1.6 万トークン）であり，総額は 0.01 ドル（約 1.5 円）未満となった．

手動データ補完. 最終的なテストセットの目標数である 1,000 サンプルに対し，自動収集のみでは不足した 68 サンプルについては手動での収集を行った．イベントごとのサンプル数バランスを考慮しつつ実施し，これには 3.5 時間を要した．どの音響イベントが不足したかの詳細は Appendix2. を参照されたい．

以上の結果から，提案手法は評価用データセットを構築する上で，時間的・金銭的コストの双方において高い効率性を有していることが示された．

4.2.3 提案データセットの特徴量分布

提案したデータセットの音響特徴量分布を可視化した結果を図 2 に示す．可視化には t-SNE を用い，各点は各サンプルの特徴量ベクトルを 2 次元に圧縮したものを表している．図より，各音響イベントクラスが特徴空間上でクラスごとに一定のまとまりを持っていることが確認できる．これは，提案手法によって収集されたデータが，音響的な一貫性を有していることを示唆している．

4.2.4 既存データセットとの分布比較

分布の可視化による定性評価

提案データセットと既存データセット AudioCaps の分布の差異を定性的に評価するため，サンプル数の多い上位 5 つのイベントラベル（“Male speech”, “Applause”, “Female speech”, “Sneeze”, “Typing”）に着目し，両データセットを重ねて可視化した結果を図 3 に示す．

（注6）：https://y-okamoto.sakura.ne.jp/XACLE Challenge/2025/baseline_model/trained_baseline_model.zip

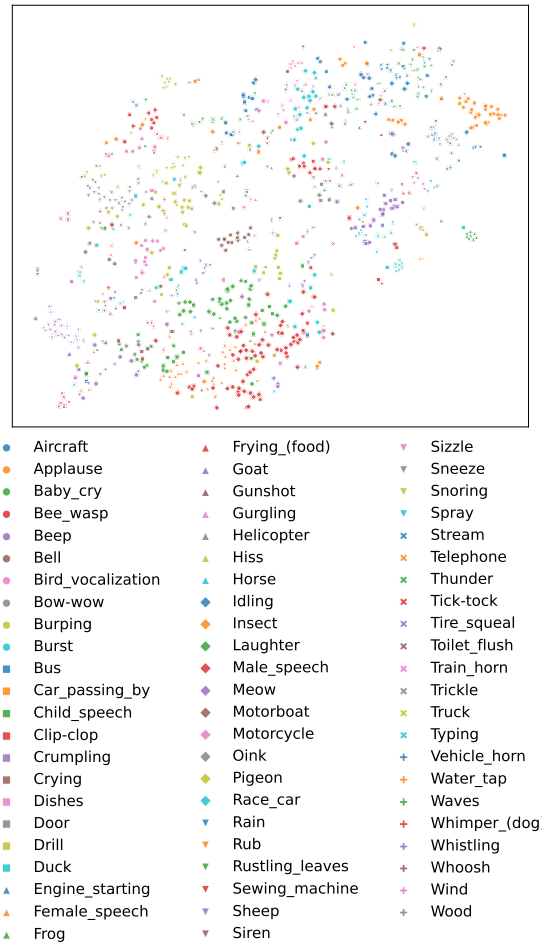

図 2　提案データセットにおける特徴量分布

表 1　各指標における提案データセットと既存データセットの分布間距離

FAD	FD	KL(Ref ‖ Tgt)	KL(Tgt ‖ Ref)
3.1711	18.9742	13.4355	14.8384

図より，同一のイベントラベルにおいて，両データセットの分布に広範な重なりが見られることが確認できる．これは，提案手法が半自動構築でありながら，既存データセットと類似した音響的特徴を持つデータを収集できていることを示唆している．すなわち，既存のベンチマークと同等のドメイン特性を維持したデータセット構築ができていることが定性的に確認された．

分布間距離に基づく定量評価

提案データセットと既存データセット AudioCaps 間の音響特徴量分布の類似性を定量的に評価するため，FD，FAD，および KL ダイバージェンスを算出した．結果を表 1 に示す．FAD および FD の値は，提案データセットが既存データセットと一定の距離（差異）を持っていることを示している．これは，提案データセットが既存データセットの単純な複製ではないと言える．また，イベント・音質の観点で同じドメインであるが，異なる評価用データセットとして機能する可能性を裏付けている．

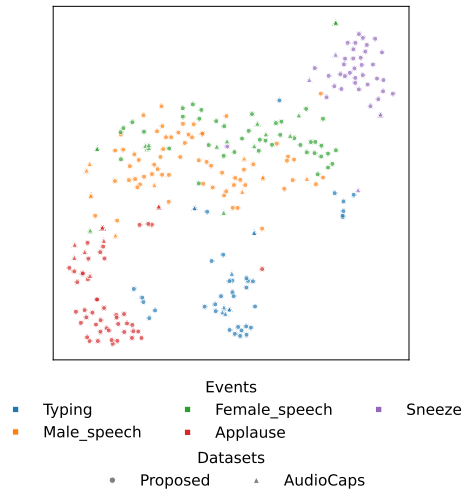

図 3　サンプル数の多い 5 イベントにおける提案データセットと既存データセットの分布比較

表 2　XACLE Challenge 2026 validation set（val_{ref}）、XACLE Challenge 2026 test set の自然音（test_{ref}）および合成音（test_{gen}）に対する RELATE モデルの性能評価

Setting	SRCC ↑	LCC ↑	KTAU ↑	MSE ↓	N sample
val_{ref}	0.177	0.187	0.123	4.385	1000
test_{ref}	0.110	0.116	0.075	3.177	1000
test_{gen}	0.270	0.259	0.184	5.699	1000

4.2.5　各データセットごとの RELATE モデルの評価比較

提案データセットが実際のモデル評価において有用であるかを検証するため，RELATE モデルを用いた評価スコアの比較を行った．表 2 に，XACLE Challenge 2026 dataset validation set（val_{ref}），XACLE Challenge 2026 dataset test set の自然音（test_{ref}），および同 test set の合成音（test_{gen}）に対する相関係数（SRCC, LCC, KTAU）および平均二乗誤差（MSE）の結果を示す．

表 2 の結果を見ると，AudioCaps 由来の自然音である validation set（val_{ref}）と，提案手法により収集された test set の自然音（test_{ref}）の間では，相関係数（SRCC, LCC, KTAU）に大きな乖離はなく，SRCC で 0.11〜0.17 程度と比較的近い値で推移しており，自然音に対する評価傾向は類似していることが確認できる．このことは，提案手法を用いて半自動的に収集されたデータ（test_{ref}）が，人手による厳密なアノテーションを経て構築された既存データセット（val_{ref}）と比較しても，同等の音響的な複雑さやドメイン特性を保持できていることを示唆している．すなわち，低コストな構築手法であっても，ベンチマークとして遜色のない品質が確保されているといえる．

これに対し，合成音データである test_{gen} については，SRCC が 0.270 と，他の自然音データ群（val_{ref}, test_{ref}）と比較して顕著に高い値を示した．自然音同士のスコアが低水準で安定しているのに対し，合成音のみが高い相関を示したこの結果は，RELATE モデルが，背景雑音や録音環境が多様な自然音よりも，テキストプロンプトの特徴が純粋に反映された合成音に対して高い感度を持つこと，あるいは合成音の方がモデルにとっ

て識別しやすい特性を持っている可能性を表している．

5. ま と め

本稿では，環境音評価データセットの半自動構築手法を提案し，その有効性を検証した．提案手法は，公開期間に基づくフィルタリング，LLM と環境音認識モデルを組み合わせた適応的クエリ更新および自動スクリーニングにより，学習データとの意図しない重複を回避しつつ，低コストで多様な環境音データを収集することを可能にした．実験的評価の結果，人手による作業を大幅に削減しながらも，実用的な精度でデータを収集できることが確認された．また，構築されたデータセットは既存データセットと比較して異なる分布特性を持ち，評価用データセットとして独自の価値を有することが示された．今後の課題として，より多様な音響イベントへの対応や，自動スクリーニング精度のさらなる向上が挙げられる．

謝辞： 本研究の一部は，JST 創発的研究支援事業 JPMJFR226V，JSPS 科研費 23K16908，24K23880，25K21221，テレコム先端技術研究支援センター 研究費助成，JST ムーンショット型研究開発事業 JPMJMS2011 の助成を受けたものです．

文 献

[1] H. Liu et al., “AudioLDM: Text-to-audio generation with latent diffusion models,” in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023, pp. 21 450–21 474.

[2] S. Chen et al., “BEATs: Audio pre-training with acoustic tokenizers,” in *Proceedings of the International Conference on Machine Learning (ICML)*, vol. 202, 2023, pp. 5178–5193.

[3] F. Kreuk et al., “AudioGen: Textually guided audio generation,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[4] D. Ghosal et al., “Text-to-audio generation using instruction-tuned llm and latent diffusion model,” 2023.

[5] Q. Kong et al., “Panns: Large-scale pretrained audio neural networks for audio pattern recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2019.

[6] C. D. Kim et al., “Audiocaps: Generating captions for audios in the wild,” in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[7] S. Lee et al., “ACAV100M: Automatic curation of large-scale datasets for audio-visual video representation learning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 274–10 284.

[8] A. Miech et al., “Howto100m: Learning a text-video embedding by watching hundred million narrated video clips,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 2630–2640.

[9] I. Martin-Morato, A. Mesaros, “What is the ground truth? reliability of multi-annotator data for audio tagging,” in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2021, pp. 46–50.

[10] B. Barz, J. Denzler, “Do we train on test data? purging cifar of near-duplicates,” *Journal of Imaging*, vol. 6, 2019.

[11] Y. Okamoto et al., “Xacle challenge 2026: The first x-to-audio alignment challenge,” UTokyo Repository, 2026.

[12] J. F. Gemmeke et al., “Audio set: An ontology and human-labeled dataset for audio events,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[13] K. J. Piczak, “Esc: Dataset for environmental sound classification,” in *Proceedings of the 23rd ACM International Conference on Multimedia*, 2015.

[14] J. Salamon et al., “A dataset and taxonomy for urban sound research,” in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014.

[15] K. Drossos et al., “Clotho: An audio captioning dataset,” 2019.

[16] X. Mei et al., “Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3339–3354, 2023.

[17] H. Chen et al., “Vggsound: A large-scale audio-visual dataset,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[18] E. Fonseca et al., “FSD50K: An open dataset of human-labeled sound events,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[19] B. Elizalde et al., “CLAP: Learning audio concepts from natural language supervision,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] Y. Gong et al., “AST: Audio spectrogram transformer,” in *Proceedings of Interspeech*, 2021, pp. 571–575.

[21] A. Radford et al., “Robust speech recognition via large-scale weak supervision,” 2022.

[22] H. Munakata et al., “Language-based audio moment retrieval,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[23] A. Vaswani et al., “Attention is all you need,” 2017.

[24] W. Chen et al., “EAT: Self-supervised pre-training with efficient audio transformer,” in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[25] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[26] M. Heusel et al., “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[27] K. Kilgour et al., “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms,” in *Proceedings of Interspeech*, 2019, pp. 2350–2354.

[28] Y. Kanamori et al., “Relate: Subjective evaluation dataset for automatic evaluation of relevance between text and audio,” in *Proceedings of Interspeech*, 2025, pp. 3155–3159.

付 録

1. キャプションおよび主観評価値の収集

XACLE Challenge 2026 test set として使用するため，収集プロセスによって得られた音響データに対するキャプションおよび主観評価値の付与を行なった．この一連の作業にはクラウドソーシングサービスを利用した．作業者には，収集された音データおよびその元となる動画を提示し，以下の 2 点のアノテーションを依頼した．

（1） **キャプション記述**：提示された音響イベントの内容を適切に表現する説明文の作成．1 サンプルにつき 1 件のキャプションを収集した．

（2） **関連度評価**：提示された音データと，記述されたキャプションとの間の意味的な関連度（relevance score; REL）の主観評価．各サンプルに対し 8 名の評価者がスコアを付与し，その平均値を当該サンプルの REL スコアとして採用した．

2. データセット構築の各過程におけるカテゴリごとのサンプル数

提案手法によるデータセット構築の各段階におけるサンプル数の推移および通過率の詳細を表 A· 1 に示す．表中の各列はそれぞれ，Web からのダウンロード総数（Downloads），自動フィ

ルタリング通過数（Auto True），人手による検証通過数（Human True），不足分を補うための追加収集数（Add），および最終的なテストセットのサンプル数（Final）を表している．また，自動処理および人手処理における通過率をそれぞれ R_{auto} (Auto True / Downloads)，R_{human} (Human True / Auto True) として算出した．

2.1 自動フィルタリングによる選別 (R_{auto})

全体として，R_{auto} は平均約 3.8% と低い値にとどまった．これは，テキスト検索のみで収集された Web 上の音響データには，タグやファイル名が一致していても実際の内容が異なるケースが多く含まれることを示唆している．特に “Burst” (0.54%) や “Wind” (0.28%) といった抽象的あるいは環境依存性の高いクラスでは，意図しない音源が混入しやすい．反対に “Male speech” (35%),“Female speech”(42%), “Child speech”(29%) のような人間の声は高い割合で正と判定されている．

2.2 人手による検証 (R_{human})

自動フィルタリングを通過した後の人手検証における通過率 R_{human} は全体で約 64% であった．“Bell” や “Vehicle_horn” では 100% の高い適合率を示したが，“Crying” (10%) や “Whoosh” (8.9%) など一部のクラスでは，イベントラベルと音響データが合致しない，あるいは他の音響イベントの割合が高く含まれていたことが確認された．モデルの判定と人間による判定の乖離を小さくすることが今後の課題である．

2.3 人手による収集 (Add)

“Human_Add” の列は自動収集ではサンプル数が 5 件に満たない，または少ない音響イベントに対して手動で追加したサンプル数を表している．自動収集で集まらなかった要因としては YouTube 上に対象の音響イベントを含むデータが少ないことが挙げられる．また，収集フローにおいて関連語生成の質，音データの切り出し方，自動スクリーニングの精度などの複数の要因が考えられる．

表 A· 1　各構築段階におけるサンプル数の統計．

Event	Down.	Auto True	Human True	Add	Final	R_{auto}	R_{human}
Aircraft	756	28	23	0	23	0.0370	0.82
Applause	585	59	45	0	45	0.1009	0.76
Baby_cry	530	15	10	3	13	0.0283	0.67
Bee_wasp	1042	63	27	0	27	0.0605	0.43
Beep	377	10	8	1	9	0.0265	0.80
Bell	446	19	19	0	19	0.0426	1.00
Bird_vocalization	161	30	30	0	30	0.1863	1.00
Bow-wow	755	22	11	2	13	0.0291	0.50
Burping	552	25	23	0	23	0.0453	0.92
Burst	736	4	4	1	5	0.0054	1.00
Bus	443	5	5	3	8	0.0113	1.00
Car_passing_by	343	3	2	3	5	0.0087	0.67
Child_speech	168	49	21	0	21	0.2917	0.43
Clip-clop	743	9	6	0	6	0.0121	0.67
Crumpling	290	26	25	0	25	0.0897	0.96
Crying	722	21	2	3	5	0.0291	0.10
Dishes	828	13	2	3	5	0.0157	0.15
Door	978	12	5	1	6	0.0123	0.42
Drill	954	25	11	0	11	0.0262	0.44
Duck	688	26	13	0	13	0.0378	0.50
Engine_starting	664	8	5	0	5	0.0120	0.62
Female_speech	100	42	42	0	42	0.4200	1.00
Frog	447	23	20	0	20	0.0515	0.87
Frying_(food)	741	30	9	0	9	0.0405	0.30
Goat	838	11	2	3	5	0.0131	0.18
Gunshot	268	6	4	1	5	0.0224	0.67
Gurgling	816	2	2	5	7	0.0025	1.00
Helicopter	445	23	10	0	10	0.0517	0.43
Hiss	659	4	3	3	6	0.0061	0.75
Horse	1061	20	4	1	5	0.0189	0.20
Idling	625	17	14	0	14	0.0272	0.82
Insect	559	15	7	0	7	0.0268	0.47
Laughter	454	33	24	0	24	0.0727	0.73
Male_speech	181	64	64	0	64	0.3536	1.00
Meow	687	25	10	0	10	0.0364	0.40
Motorboat	681	10	3	2	5	0.0147	0.30
Motorcycle	584	27	11	0	11	0.0462	0.41
Oink	639	10	2	3	5	0.0156	0.20
Pigeon	355	28	16	0	16	0.0789	0.57
Race_car	480	38	20	0	20	0.0792	0.53
Rain	331	41	14	0	14	0.1239	0.34
Rub	613	3	1	4	5	0.0049	0.33
Rustling_leaves	512	4	3	4	7	0.0078	0.75
Sewing_machine	636	11	5	1	6	0.0173	0.45
Sheep	506	37	21	0	21	0.0731	0.57
Siren	465	14	7	0	7	0.0301	0.50
Sizzle	817	6	5	2	7	0.0073	0.83
Sneeze	693	50	36	0	36	0.0722	0.72
Snoring	331	35	29	0	29	0.1057	0.83
Spray	974	24	4	1	5	0.0246	0.17
Stream	451	13	13	2	15	0.0288	1.00
Telephone	949	1	1	4	5	0.0011	1.00
Thunder	220	39	19	0	19	0.1773	0.49
Tick-tock	952	20	19	0	19	0.0210	0.95
Tire_squeal	391	8	6	0	6	0.0205	0.75
Toilet_flush	508	19	13	0	13	0.0374	0.68
Train_horn	391	29	28	0	28	0.0742	0.97
Trickle	486	17	14	0	14	0.0350	0.82
Truck	410	14	7	0	7	0.0341	0.50
Typing	433	42	36	0	36	0.0970	0.86
Vehicle_horn	511	11	11	0	11	0.0215	1.00
Water_tap	711	4	2	3	5	0.0056	0.50
Waves	205	30	26	0	26	0.1463	0.87
Whimper_(dog)	471	9	9	0	9	0.0191	1.00
Whistling	178	37	33	0	33	0.2079	0.89
Whoosh	758	22	2	3	5	0.0290	0.09
Wind	711	2	1	4	5	0.0028	0.50
Wood	892	15	2	3	5	0.0168	0.13
Total	38887	1457	931	69	1000	0.0375	0.64