環境音埋め込みベクトル系列の類似度に基づく環境音合成の自動評価

岸 秀[1]　阪井 瞭介[1]　高道 慎之介[1,2]　金森 勇介[2]　岡本 悠希[2]

概要：環境音合成タスクにおいて，合成音に対する主観評価は重要な評価基準だが，その実施には時間的・金銭的なコストがかかる．そのため，メルケプストラム歪みや CLAPScore などの客観評価指標も用いられるが，その客観指標値と主観評価値の相関は弱い．本稿では，環境音合成タスクにおける客観評価指標として AudioBERTScore を提案する．これは，合成音と参照音それぞれの埋め込み表現に基づいた系列間の類似度を計算する手法である．提案手法では，従来の BERTScore を流用した最大値ノルムに加え，環境音の非局在性を反映した p-ノルムに基づく．実験的評価の結果，提案手法による評価値は，従来のメルケプストラム歪みなどより，主観評価値との相関が高いことを示す．

1. はじめに

テキスト環境音合成（text-to-audio; TTA）モデルは「小さな犬が鳴いている」のようなテキストを入力にとり環境音を生成する深層学習モデルである．代表的なものに AudioLDM2 [1] や AudioLCM [2] がある．TTA で合成された環境音はアニメやゲームなどのメディアコンテンツ作成に使用されており [3]，コンテンツの状況設定や人物の心情の提示などにおいて重要な役割を果たす [4]．TTA モデルの性能はその合成音により評価され，種々の評価法のうち人間による主観評価は最も重要な手法として扱われる [5]．実際，DCASE 2024 Challenge（Detection and Classification of Acoustic Scenes and Events）[*1] のような国際的コンペティションの TTA の最終的な評価は主観評価値に基づく．

合成音の主観評価は，全体的な音質 (overall quality; OVL) [6] や入力テキストとの合致度 (relevance to the input text; REL) [5] の観点で実施されることが多いが，実施にかかる時間的・金銭的コストが高い．そのため，メルケプストラム歪み（mel cepstral distortion; MCD） [7] や CLAPScore [8] などの客観評価指標が提案されている．客観評価指標の性能を測る上で主観評価との整合を見ることが重要である [5] が，既存の客観評価指標の値と主観評価値の相関は低い [9].

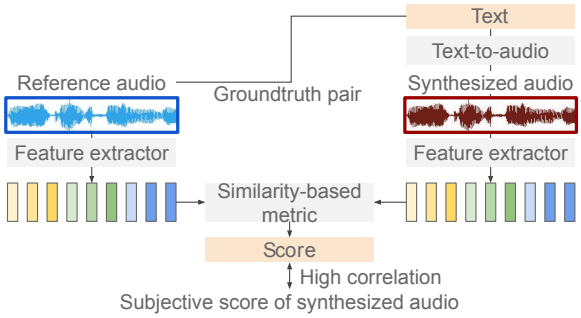

図 1　提案する客観評価指標の概要

そこで本研究では，図 1 に示す新たな客観評価指標 AudioBERTScore を提案する．この手法は，合成音と参照音（reference audio）を使用する．環境音基盤モデルを用いて，各音から埋め込みベクトル系列を抽出したのち，系列間の類似度に基づいて評価値を推定する．この時，環境音基盤モデルとして BYOL-A（Bootstrap Your Own Latent for Audio） [10] や AST（Audio Spectrogram Transformer） [11] などで比較実験を行うことで AudioBERTScore に最適な特徴抽出器を調査する．実験的評価において合成音に対する主観評価値との相関を調査した結果，合成音と参照音を用いる客観評価指標の中で，提案手法による客観評価値が主観評価値と最も強く相関することを示す．提案手法は，オープンソースのツールキットとして公開している[*2].

[1] 慶應義塾大学
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, Kanagawa 223-8522, Japan.
[2] 東京大学
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan.

[*1] https://dcase.community/challenge2024/task-sound-scene-synthesis

[*2] https://github.com/lourson1091/audiobertscore

表 1　自動評価指標の分類.

Method	Training	Reference	Pretrained model
MCD	No	Yes (audio)	No
WARP-Q	No	Yes (audio)	No
FAD	No	Yes (audio)	Yes (audio)
AudioBERTScore (ours)	No	Yes (audio)	Yes (audio)
RELATE bench	Yes	Yes (text)	Yes (text & audio)
CLAPScore	No	Yes (text)	Yes (text-audio)
PAM	No	No	Yes (text-audio)

2. 関連研究

2.1 TTA の合成音に対する評価指標

TTA の合成音はしばしば，主観評価と客観評価（自動評価）の両方で評価される．主観評価の代表的な観点として，OVL と REL がある [12]．前者は合成音自体の品質を，後者は TTA の入力テキストとの対応を評価者に回答させる．主観評価は主観品質を定量化する有用な手段だが，実施に係る時間的・金銭的コストが高い．

そこで，主観評価値に相関する値を推定可能な自動評価が研究されている．本稿では種々の自動評価を以下の観点で分類する．

- **学習あり／なし**：主観評価値を用いた学習を必要とするか否か．主観評価値を予測する直接的な方法は，テキストおよび合成音と主観評価値の対データを用意して，合成音から主観評価値を予測する機械学習モデルを学習することである．本観点では，そのような対データおよび学習の有無で分別する．
- **参照あり／なし**：評価値推定の際に用いる参照データの有無，および，用いる場合の種類．合成音に加え，合成に使用したテキストを用いるもの，groundtruth となる音を用いるもの，あるいは合成音のみを用いるものに分別する．
- **事前学習モデルの使用あり／なし**：事前学習モデルを推定に用いるか否か，および，用いる場合の種類．自己教師あり学習のように，主観評価値を用いないが，事前の学習を必要とするもの．事前学習を用いる場合は，事前学習に使用するデータの種類で分別する．

これらに基づいて既存手法を分類した結果を表 1 に示し，その詳細を以下に示す．

- **MCD, WARP-Q：** 学習なしで参照音を用いる手法である．メルケプストラム歪み (MCD) は，合成音と参照音の間のメルケプストラム距離であり [13]，WARP-Q は各音を短時間セグメントに分割したときのメル周波数ケプストラム距離 [14] である．これらの値は参照音から計算可能であるため，TTA の入力テキストの言語に依存しない強みがあるが，4 節に示すように主観評価値との相関が低い．
- **RELATE bench：** 合成音声品質値の教師あり自動予測の成功 [15] を受け提案されたのが，RELATE ベンチマークシステムである [16]．REL に関する主観評価値データセットを新たに作成し，合成音とテキストのペアから主観評価値を予測するモデルを教師あり学習する．教師あり学習により高い相関を期待できるが，大規模なデータセットの用意が必須である．
- **PAM：** PAM (prompting audio-language models for audio quality assessment) [17] は，テキストと音の対照学習モデルを用いて，音質に言及するテキストプロンプトと合成音から評価値を推定する．PAM は参照不要である強みを持つが，テキストと音の対データを用いた事前学習が必要であるため，そのような対データの存在する言語でしか使用できない．

本研究で提案する AudioBERTScore は，MCD や WARP-Q と同じく，学習なしで参照音を用いる手法である．特定の言語に依存しない強みを保ちつつ，環境音基盤モデルによる特徴量抽出，距離（類似度）計算の改善により，主観評価値との高い相関を目指す．

2.2 SpeechBERTScore

主観評価値と相関する自動評価値の算出は，TTA に限らずテキスト音声合成においても需要がある．この客観評価指標の例として SpeechBERTSocre [18] がある．SpeechBERTScore は，自然言語処理分野のテキスト生成タスクにおいて提案された BERTScore [19] に基づいて設計されている．BERTScore は，生成文と参照文それぞれから文脈埋め込みベクトル系列を計算し，その系列間の類似度を評価値とする．自己教師あり学習モデル Bidirectional Encoder Representations from Transformers (BERT) [20] を用いた埋め込みベクトル計算，および，系列間の強制アライメントを意識した類似度計算により，高い相関の自動評価を実現している．SpeechBERTScore はこの内，自己教師あり学習 (self-supervised leaning; SSL) モデルを音声に特化したモデル（例えば，HuBERT [21]，WavLM [22]）に置換することで，合成音声の自動評価に成功している．

本稿で提案する AudioBERTScore もこの流れを汲む．具体的には，埋め込みベクトルの算出に環境音基盤モデルを利用する．本稿ではさらに，環境音における系列アライメント問題を再考し，新たな類似度計算を提案する．

3. 提案する客観評価指標

本節では提案する TTA の客観評価指標である AudioBERTScore について述べる．

3.1 特徴量抽出と類似度行列

3.1.1 特徴量抽出

AudioBERTScore でスコアを計算するには合成音，参照音のそれぞれの埋め込みベクトル系列を得る．まず，合成

音の波形を $S = (\boldsymbol{s}_t \in \mathbb{R} \mid t = 1, \cdots, T_{\text{gen}})$，参照音の波形を $R = (\boldsymbol{r}_t \in \mathbb{R} \mid t = 1, \cdots, T_{\text{ref}})$ と表現する．ここで，T_{gen} と T_{ref} は一致する必要はない．$\tilde{S} = (\tilde{\boldsymbol{s}}_n \in \mathbb{R}^D \mid n = 1, \cdots, L_{\text{gen}})$ および $\tilde{R} = (\tilde{\boldsymbol{r}}_n \in \mathbb{R}^D \mid n = 1, \cdots, L_{\text{ref}})$ は，それぞれ S と R から特徴抽出器を用いて抽出された埋め込みベクトル系列を示し，以下の式で表される．

$$\tilde{S} = \text{Encoder}(S; \theta) \tag{1}$$

$$\tilde{R} = \text{Encoder}(R; \theta) \tag{2}$$

ここで θ は，事前学習済みの特徴抽出器のパラメータを示す．L_{gen} および L_{ref} は，それぞれ T_{gen} および T_{ref} と，特徴抽出器のフレームレートによって決定される．

特徴抽出器の選択には任意性がある．例えば，Transformer[23] に基づく自己教師あり学習モデルを利用する場合，浅い層の特徴量は音響的，深い層の特徴量は文脈的特徴を有すると期待される．

3.1.2 類似度行列

3.1.1 節で得られた埋め込みベクトル系列 $\tilde{S}$ および $\tilde{R}$ に対して，各ベクトル間の類似度を計算することで，合成音と参照音との特徴量の類似度を行列表現する．類似度行列 $M \in \mathbb{R}^{L_{\text{gen}} \times L_{\text{ref}}}$ の (i, j) 番目の要素 M_{ij} は以下のコサイン類似度で与えられる．

$$M_{ij} = \text{sim}(\tilde{\boldsymbol{s}}_i, \tilde{\boldsymbol{r}}_j) = \frac{\tilde{\boldsymbol{s}}_i \cdot \tilde{\boldsymbol{r}}_j}{\|\tilde{\boldsymbol{s}}_i\| \cdot \|\tilde{\boldsymbol{r}}_j\|} \tag{3}$$

3.2 スコア計算

類似度行列に基づいてスコアを計算する．まず，BERTScore と SpeechBERTScore で利用されている，最大値ノルムに基づく計算法を AudioBERTScore に転用する．次に，自然言語，音声，環境音の間の局在性の違いに着目して，p-ノルムに基づく計算法を提案する．計算の概要は図 2 の通りである．

3.2.1 最大値に基づく計算

類似度行列から，適合率 (precision)，再現率 (recall)，F1 スコアを計算する．まず，適合率は，合成音埋め込みの各フレームについて類似度の最大値を取得し，それらを平均する．すなわち，合成音に含まれる要素を参照音がどの程度含むかを意味する．最大値に基づく適合率 $\text{precision}_{\text{max}}$ は下式で計算される．

$$\text{precision}_{\text{max}} = \frac{1}{L_{\text{gen}}} \sum_{i=1}^{L_{\text{gen}}} \max_{j=1,\ldots,L_{\text{ref}}} M_{ij} \tag{4}$$

次に，再現率は，適合率計算における合成音と参照音を入れ替えて計算され，参照音に含まれる要素を合成音がどの程度含むかを表す．最大値に基づく再現率 $\text{recall}_{\text{max}}$ は下式で計算される．

$$\text{recall}_{\text{max}} = \frac{1}{L_{\text{ref}}} \sum_{j=1}^{L_{\text{ref}}} \max_{i=1,\ldots,L_{\text{gen}}} M_{ij} \tag{5}$$

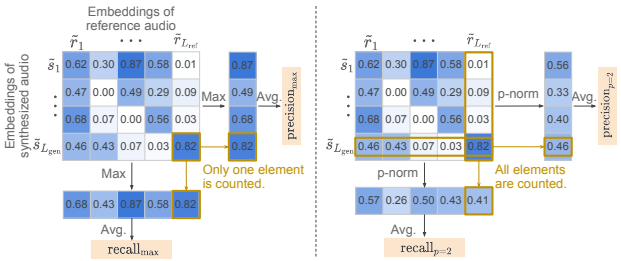

図 2 提案法における類似度の計算．左：BERTScore に由来する，局在性を前提とする最大値に基づく手法．右：環境音の非局在性を反映する p-ノルムに基づく手法．$\text{precision}_*, \text{recall}_*$ はそれぞれ，類似度行列の行方向あるいは列方向ごとに値を集計した後，その平均値を計算する．集計の際，最大値に基づく手法（左）では特定の要素の値のみを抽出するが，p-ノルムに基づく手法（右）では該当する要素全体を用いて計算する．

この適合率と再現率の計算は図 2 左に図示されている．

そしてこれらの適合率と再現率の調和平均として $F1$ スコア F1_{max} を以下のように計算する．

$$\text{F1}_{\text{max}} = 2 \times \frac{\text{precision}_{\text{max}} \times \text{recall}_{\text{max}}}{\text{precision}_{\text{max}} + \text{recall}_{\text{max}}} \tag{6}$$

BERTScore と SpeechBERTScore を転用したこれらのスコアリング手法は，∞ ノルム（最大値ノルム）に基づいている．すなわち，「類似度の高いベクトルは時間的に局所的に現れる」という局在性の仮定を置いている．この仮定は，自然言語や音声を評価対象とする場合には概ね妥当である．自然言語では特定のフレーズが，音声では分節的特徴が局所的に出現するためである．

一方，環境音ではこの仮定が必ずしも成立しない．たとえば銃声のように瞬時に生起する音では，対応する埋め込みベクトルが短い時間区間に集中するため，局在性の仮定は有効である．しかし渓流のせせらぎのように非構造的で持続的な音では，埋め込みが時間全体にわたって広がり（すなわち非局在的になり），前提が崩れる．こうした音も適切に評価するには，非局在性を捉えられるスコア計算手法の導入が有効であると考えられる．

3.2.2 p-ノルムに基づく計算

3.2.1 節の最大値ノルムを p-ノルムに置換した下式を新たにスコアとして用いる．

$$\text{precision}_p = \frac{1}{L_{\text{gen}}} \sum_{i=1}^{L_{\text{gen}}} \left(\frac{1}{L_{\text{ref}}} \sum_{j=1}^{L_{\text{ref}}} M_{ij}^p \right)^{1/p} \tag{7}$$

$$\text{recall}_p = \frac{1}{L_{\text{ref}}} \sum_{j=1}^{L_{\text{ref}}} \left(\frac{1}{L_{\text{gen}}} \sum_{i=1}^{L_{\text{gen}}} M_{ij}^p \right)^{1/p} \tag{8}$$

図 2 右に示すように，precision_p と recall_p は p-ノルムを用いて定義される．$p = 1$ の場合，両指標は単純平均と一致し，合成音と参照音の埋め込みベクトルが時間全体でどれだけ一致しているか（すなわち非局在的な類似度）を測定する．p を大きくするにつれて，計算は局所的なピークに

重み付けされ，局在性を捉える性質が強まる．極限 $p \to \infty$ では，指標は ∞-ノルム（最大値ノルム）と同値になる．

さらに，最大値に基づく計算と p-ノルムに基づく計算を内分した下式を導入することで，局所的・非局所的の双方を同時に評価できるようする．

$$\mathrm{precision}_{\lambda,p} = \lambda \cdot \mathrm{precision}_{\max} + (1-\lambda) \cdot \mathrm{precision}_{p} \tag{9}$$

$$\mathrm{recall}_{\lambda,p} = \lambda \cdot \mathrm{recall}_{\max} + (1-\lambda) \cdot \mathrm{recall}_{p} \tag{10}$$

$$\mathrm{F1}_{\lambda,p} = 2 \times \frac{\mathrm{precision}_{\lambda,p} \times \mathrm{recall}_{\lambda,p}}{\mathrm{precision}_{\lambda,p} + \mathrm{recall}_{\lambda,p}} \tag{11}$$

$0 \leq \lambda \leq 1$ は内分のハイパーパラメータであり，$\lambda = 1$ あるいは $p \to \infty$ で最大値ノルムに一致する．

4. 実験的評価

提案手法の有効性を評価するために，提案手法による自動評価値と，人間による主観評価値の相関を計算した．

4.1 実験条件

4.1.1 データセット

評価のために，合成音と参照音の対，および，合成音に対する主観評価値からなるデータセットを使用する．本研究では PAM [17] のテストセットを使用した．当該セットには，AudioCaps [24] からランダムに抽出した自然音 (参照音) と英語キャプションの 100 対と，そのキャプションから，後述する 4 つの TTA モデルを用いて作成した合成音 400 サンプルが含まれる．参照音・合成音それぞれのサンプルに，テキストと音の関連度を評価した REL，および音の品質を評価した OVL それぞれの 5 段階 MOS 値が与えられている．各 MOS 値は 10 人の異なる評価者によるスコアの平均である．本研究ではこのデータセットのうち，参照音に対する REL 主観評価値が 3.5 に満たない参照音 17 サンプル，および対応する合成音 17×4 サンプルを評価から除外した．これは，REL 主観評価値に対する提案手法の評価値を正確に評価するためである．REL 主観評価値はテキストと音の関連度を表す値であるのに対し，提案手法は参照音と合成音の対からこの値に相関する自動評価値を推定する．故にその評価には，参照音とテキストが強く関連している前提がある．そのため，前述のサンプルを除外した．各サンプルの時間長は，合成音 5 sec，参照音 10 sec である．

使用した合成モデルは MelDiffusion, AudioLDM2[*3], AudioLDM-Large[*4][25], AudioGen-base[*5][26] の 4 つであった．モデルの詳細は PAM [17] を参照されたい．サンプリング周波数は参照音 44.1 kHz，合成音 22.05 kHz および 16 kHz であり，後述する特徴抽出器に入力する際に全て 16 kHz にダウンサンプリングした．

4.1.2 特徴抽出器

提案手法の特徴抽出器として，以下の 3 つの学習済みモデルを用意した．

- **BYOL-A [10]**[*6]：畳み込みニューラルネットワーク (convolutional neural network; CNN) [27] に基づくモデル．最新の v2 を使用した．畳み込みニューラルネットワークの出力をチャネルと周波数を 1 次元に平坦化したフレームレベルの埋め込み (“local”)，それを多層パーセプトロンに通して得る埋め込み (“global”)，これら二つの埋め込みを特徴次元方向に結合したフレーム–全体埋め込み (“local+global”) を使用する．ここでフレーム方向のプーリングは行わない．
- **ATST-Frame [28]**[*7]：13 層の Transformer [23] に基づく，teacher-student 学習のモデル．1–13 層の各特徴量を使用する．重みは ATST-Frame-base を使用した．
- **AST [11]**[*8]：13 層の Transformer [23] に基づくモデル．画像の 1000 クラス識別タスクで Vision Transformer [29] を事前学習された重みが 527 の環境音クラスへの識別タスクにより fine-tuning されている．1–13 層の各特徴量を使用する．fine-tuning 済みの重みは Full AudioSet, 10 tstride, 10 fstride, with Weight Averaging (0.459 mAP) を使用した．

4.1.3 比較手法

比較手法として，提案手法と同じ条件（表 1）である学習なし，参照音使用の MCD [7]，WARP-Q [30] を使用した．同条件に FAD [31] も含まれるが，提案手法および MCD，WARP-Q がサンプル毎に計算されるのに対し，FAD は複数サンプル毎に計算されるため，比較から除外した．また，参考値として，異なる条件の評価指標である CLAPScore [8] と PAM [17] も用いた．

4.1.4 評価方法

OVL 主観評価値と REL 主観評価値のそれぞれについて，客観評価指標値との線形相関係数 (linear correlation coefficient; LCC) とスピアマンの順位相関係数 (Spearman’s rank correlation coefficient; SRCC) を計算した．テストセットの各サンプルについて値を計算し，テストセット全体で相関係数を計算した．

4.2 結果

4.2.1 特徴抽出器と適合率・再現率・F1 の比較

特徴抽出器の比較．各特徴抽出器における層ごとの出力

[*3] https://github.com/haoheliu/AudioLDM2
[*4] https://github.com/haoheliu/AudioLDM
[*5] https://github.com/facebookresearch/audiocraft
[*6] https://github.com/nttcslab/byol-a/blob/master/v2/AudioNTT2022-BYOLA-64x96d2048.pth
[*7] https://github.com/Audio-WestlakeU/audiossl/blob/main/audiossl/methods/ATST-Frame/README.md
[*8] https://github.com/YuanGongND/ast/blob/master/pretrained_models/README.md

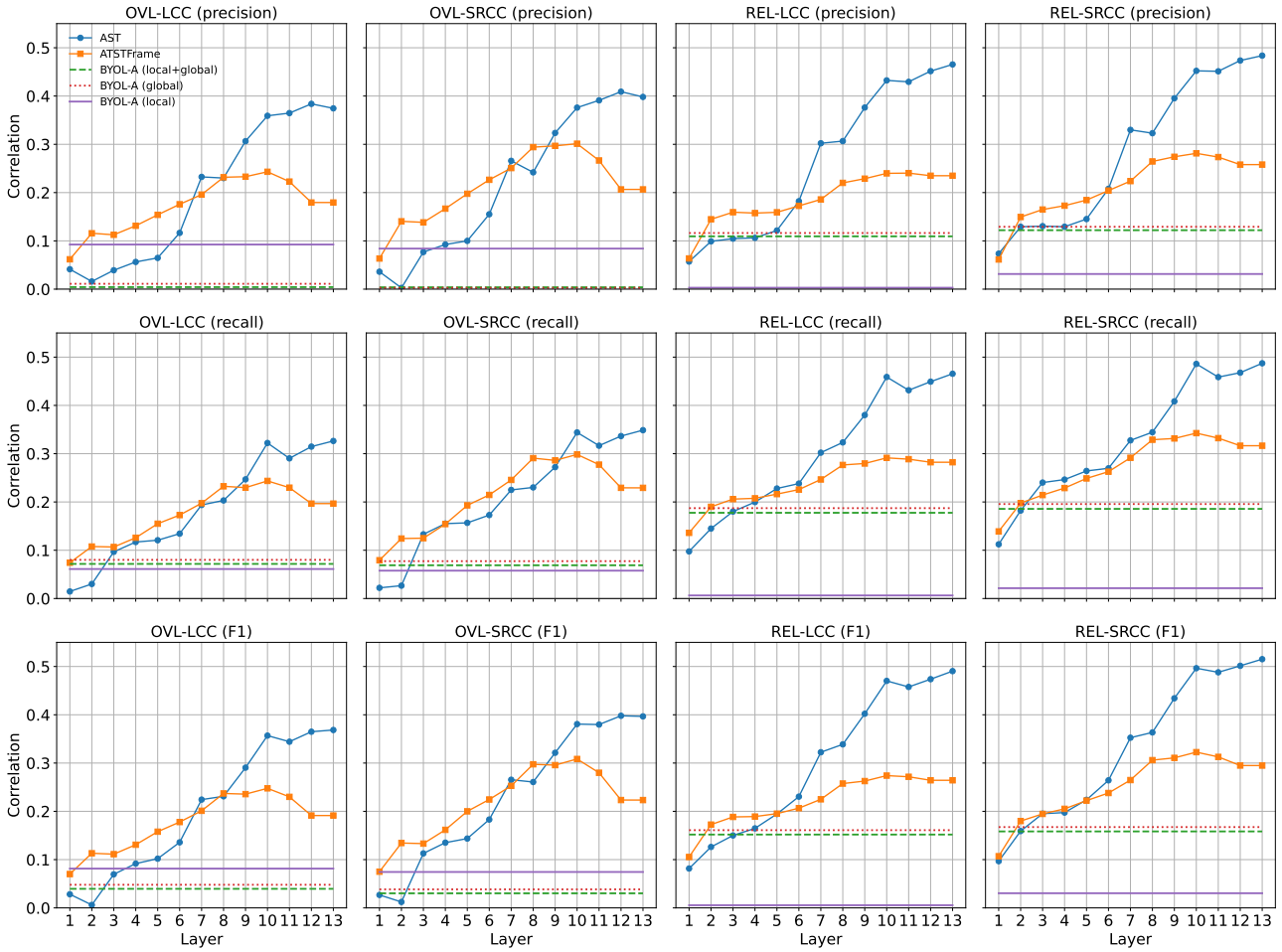

図 3 各特徴抽出器における層ごとの出力を用いて計算したスコアと主観評価値の相関比較．3 種類の特徴抽出器（AST, ATST-Frame, BYOL-A）の各層から得られる埋め込みベクトル系列を基に計算した類似度から，最大値に基づく適合率・再現率・F1 の各スコアを計算する．これらの各スコアと主観評価値（OVL, REL）の相関（LCC, SRCC）の比較．

表 2 特徴抽出器ごとの比較．最大値に基づくスコア計算と抽出する特徴は最良の設定を使用している．

	OVL		REL	
	LCC	SRCC	LCC	SRCC
AST	**0.368**	**0.397**	**0.490**	**0.515**
ATST-Frame	0.248	0.308	0.291	0.343
BYOL-A	0.080	0.077	0.187	0.196

をもとに計算したスコアと主観評価値の相関値を比較した．スコア計算には，最大値に基づく計算（式 (4)(5)(6)）を用いた．結果を図 3 に示す．AST， ATST-Frame ともに後半の層が前半の層に比べて相関が高いことがわかる．より後半の層の特徴が環境音の文脈情報を含んでおり，AudioBERTScore にとってより効果的であると言える．ただし，AST は最後の層まで比較的単調に増加するのに対し，ATST-Frame は第 11 層から減衰傾向にある．BYOL-A の OVL と REL を比較すると，OVL では local が，REL では global あるいは local+global のスコアが比較的高い．これは，入力層により近い local では音響的特徴を，それをプーリングした global では文脈情報を保持しているためだと考えられる．

相関値同士の相関の調査．AST，ATST-Frame において OVL と REL の値に着目すると，OVL との相関が高いときには REL においても相関が高い傾向が見られる．ここで，OVL と REL の関係を調べるため，図 4 に OVL-LCC と REL-LCC の相関を示す．AST，ATST-Frame 共に OVL-LCC と REL-LCC の間に高い相関が見られる．また，特徴抽出器に依らず OVL-LCC と REL-LCC の完全一致を示す対角線よりも全体的に上に位置していることから，REL-LCC の方が OVL-LCC よりも高くなる傾向があると言える．これは AudioBERTScore の設計が文脈を捉えるのに適しているからだと考える．

適合率・再現率・F1 の比較．precision と recall を比較すると，AST と ATST-Frame ともに recall が優れる傾向にある．また，AST においては，precision が第 6–7 層で，

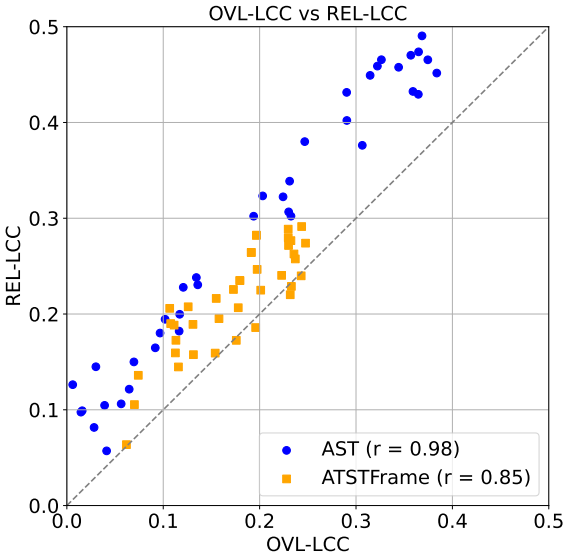

図 4 提案手法における OVL-LCC と REL-LCC の散布図．青色が AST モデル、オレンジ色が ATST-Frame モデルにおける相関である．各モデルにおける precision、recall、F1 の指標および 13 層の相関が統合されている．また，対角線（灰色点線）は OVL-LCC と REL-LCC の完全一致を示す．凡例内の r 値は各モデルの OVL-LCC と REL-LCC 間のピアソン相関係数である．

recall が第 8–10 層で急激に上昇する傾向にある．これらの理由についてはさらなる調査が必要である．

最良設定同士の比較．各特徴抽出器について最良の設定を用いたときの結果を表 2 に示す．最良設定とは，図 3 において平均的に高い相関値だった層とスコア計算の選択であり，AST は 13 層目の $\mathrm{F1}_{\max}$，ATST-Frame は 10 層目の $\mathrm{recall}_{\max}$，BYOL-A は global 層の $\mathrm{recall}_{\max}$ である．Transformer に基づく AST と ATST-Frame は，CNN に基づく BYOL-A を大きく上回ることから，Transformer による文脈情報の獲得が特徴抽出に有効だと考えられる [32]．また，AST が ATST-Frame より優れることから，両者の大きな違いである，環境音クラス識別の fine-tuning がこの際に寄与したと予想される．

4.2.2 最大値に基づく計算と p-ノルムに基づく計算の比較

p, λ の影響．p-ノルムに基づく計算を取り入れたスコア（式 (9)(10)(11)）の p, λ を変化させたときの性能を調査した．結果を図 5 に示す．4.2.1 節より，最も性能の高かった AST の 13 層目の F1 スコアを用いた．λ の値によらず $p = 100$ において最も性能が良く，特に $p = 100, \lambda = 0$ が最も性能が良い．そのため，提案手法において非局在性を捉えるスコア計算の有効性が示される．なお，$p = 100$ 近傍で性能を比較した結果，$p = 106$ が $p = 100$ よりもわずかに良い性能を示した．

負の λ の調査．図 5 のグラフの傾きに着目すると，$p = 100$ において λ が小さくなるに連れて相関は単調に高くなり，

表 3 客観評価指標ごとの評価結果．高いほど主観評価値との相関が高い．

	OVL		REL	
	LCC	SRCC	LCC	SRCC
Compared metrics				
MCD	0.004	0.008	0.029	0.05
WARP-Q	0.241	0.228	0.202	0.213
Proposed AudioBERTScore (AST, 13-th layer)				
$\mathrm{F1}_{\max}$	0.368	0.397	0.490	0.515
$\mathrm{F1}_{\lambda=0,p=106}$	0.401	0.420	0.523	0.545
$\mathrm{F1}_{\lambda=-3.5,p=106}$	**0.424**	**0.433**	**0.546**	**0.567**
Other metrics				
PAM	0.595	0.604	0.529	0.556
CLAPScore	0.337	0.323	0.487	0.475

$\lambda = 0$ においてもその傾向は続いている．このことから，内分ではなく外分になるが，さらなる改善を目的として負の λ における相関を調査した．結果を図 6 に示す．図 6 より，$\lambda = -4$ まで相関は高くなっている．しかし，$\lambda = -5$ において OVL-LCC，REL-LCC が大幅に下がっている．$\lambda = -1$ よりも負に大きくすることは p-ノルムに基づくスコアの重みが大きくなると同時に最大値に基づくスコアの絶対値も大きくなり，局在的な類似度と非局在的な類似度の両方を取り入れた結果スコアが高くなっていると考える．

4.2.3 他の客観評価指標との比較

提案手法と従来の客観評価指標の性能を比較した．結果を表 3 に示す．提案手法の性能として，4.2.1 節および 4.2.2 節後半それぞれの調査で最良のものを示している．さらにスコア計算のパラメータ p, λ は予備実験および図 6 にて最も性能の良かった $p = 106$, $\lambda = -3.5$ を使用した．また，直接的な比較対象ではないが，PAM, CLAPScore の性能も合わせて示している．

提案手法の性能は REL，OVL の両方の観点において，比較対象である WARP-Q，MCD を大きく上回ることがわかる．故に提案手法は，学習不要で参照音に基づく指標として有用であることが示される．また，提案法同士を比較すると，特徴抽出器とスコア計算を選定した $\mathrm{F1}_{\max}$，p-ノルムを導入した $\mathrm{F1}_{\lambda=0,p=106}$，ノルムの外分で得られた $\mathrm{F1}_{\lambda=-3.5,p=106}$ のそれぞれが，性能改善に貢献していることがわかる．

参考までに，提案手法の性能を PAM, CLAPScore と比較した．これらと比較しても提案手法は高い性能を発揮しており，特に REL 主観評価値とは最も高く相関している．この結果は，提案法の貢献を後押しするものである．OVL 主観評価値との相関については PAM の優位性が認められるが，本件については今後の調査予定とする．

5. おわりに

本稿では環境音埋め込みベクトル系列の類似度に基づく

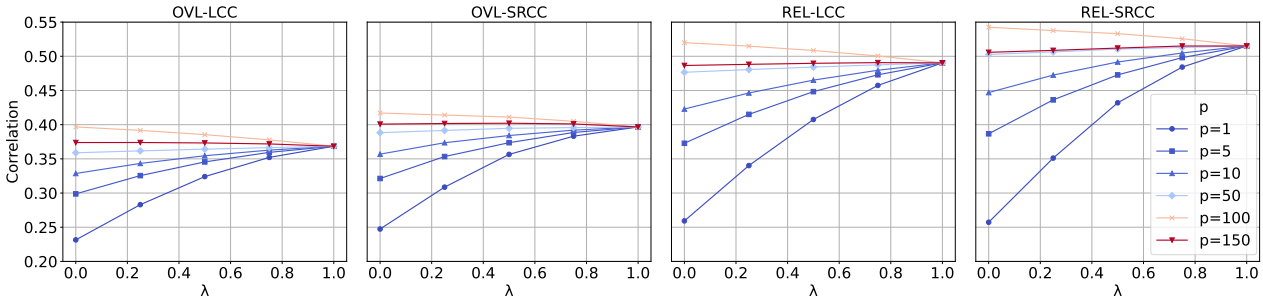

図 5 各 p, λ における p-ノルムに基づく計算法を取り入れた提案手法と主観評価値の相関.

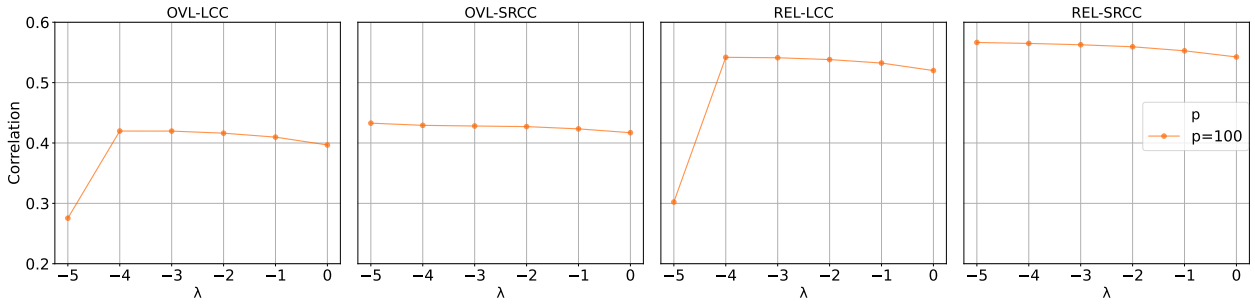

図 6 各 $p = 100$ における λ を負の範囲で変化させたときの，p-ノルムに基づく計算法を取り入れた提案手法と主観評価値の相関.

環境音合成の客観評価指標を提案した．評価結果から，教師なし・参照音ありの指標の中で最も優れる性能であることを示した．また，他条件の指標の比較でも提案法は優れた性能であることを示した．今後は，さらなる改善のための類似度計算・スコア計算手法を検討する．また，本研究を遂行する過程で，既存の学習済みモデルの事前学習データセットと，一般的に使われる評価セットの間でデータリークの疑いを発見した．これを解消するため，新たな評価データセットの策定を検討する．詳細については続報にて報告する．

謝辞 本研究の一部は，科研費 23K18474，25K21221，JST 創発的研究支援事業 JP23KJ0828，JST ムーンショット型研究開発事業 JPMJMS201 の助成を受け実施しました．

参考文献

[1] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[2] H. Liu, R. Huang, Y. Liu, H. Cao, J. Wang, X. Cheng, S. Zheng, and Z. Zhao, "AudioLCM: Text-to-audio generation with latent consistency models," *arXiv preprint arXiv:2406.00356*, 2024.

[3] T. Marrinan, P. Akram, O. Gurmessa, and A. Shishkin, "Leveraging ai to generate audio for user-generated content in video games," *arXiv preprint arXiv:2404.17018*, 2024.

[4] 掛須春希, 伊藤彰教, and 伊藤謙一郎, "日本の SF アニメにおける心情・状況演出音の演出技法の調査と制作手法分析," **画像電子学会研究会講演予稿**, vol. 16.04, pp. 239–240, 2017.

[5] Y. Okamoto, K. Imoto, S. Takamichi, T. Fukumori, and Y. Yamashita, "How should we evaluate synthesized environmental sounds," in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 307–312.

[6] J. H. L. Hansen and B. L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," *5th International Conference on Spoken Language Processing*, 1998.

[7] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1, 1993, pp. 125–128 vol.1.

[8] F. Xiao, J. Guan, Q. Zhu, X. Liu, W. Wang, S. Qi, K. Zhang, J. Sun, and W. Wang, "A reference-free metric for language-queried audio source separation using contrastive language-audio pretraining," in *Proceedings of Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop*, 2024.

[9] 高野大成, 岡本悠希, and 齋藤佑樹, "環境音合成における客観評価指標 CLAPScore の時系列評価性能の分析," in **日本音響学会 春季研究発表会講演論文集**, no. 3-Q-3, 2025, pp. 285–288.

[10] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Exploring pre-trained general-purpose audio representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 137–151, 2023.

[11] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spec-

trogram transformer," in *Proceedings of Interspeech*, 2021, pp. 571–575.

[12] K. Choi, J. Im, L. Heller, B. McFee, K. Imoto, Y. Okamoto, M. Lagrange, and S. Takamichi, "Foley sound synthesis at the dcase 2023 challenge," *arXiv preprint arXiv:2304.12521*, 2023.

[13] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, 1994, p. 359–370.

[14] S. Han and L. Zhang, "Dynamic time warping under subsequence," in *4th International Conference on Information Science, Electrical, and Automation Engineering*, vol. 12257, 2022, p. 122571X.

[15] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: Utokyo-sarulab system for voicemos challenge 2022," in *Proceedings of Interspeech*, 2022, pp. 4521–4525.

[16] 金森祐介, 岡本悠希, 高野大成, 高道慎之介, 齋藤佑樹, and 猿渡洋, "Text-to-audio における入出力関連性の自動評価に向けた主観評価データセット構築," in **情報処理学会 音声言語処理研究会**, 2025.

[17] S. Deshmukh, D. Alharthi, B. Elizalde, H. Gamper, M. A. Ismail, R. Singh, B. Raj, and H. Wang, "PAM: Prompting audio-language models for audio quality assessment," *arXiv preprint arXiv:2402.00282*, 2023.

[18] T. Saeki, S. Maiti, S. Takamichi, S. Watanabe, and H. Saruwatari, "SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics," in *Proceedings of Interspeech*, 2024, pp. 4943–4947.

[19] T. Zhang*, V. Kishore*, F. Wu*, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with bert," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[21] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, p. 3451–3460, Oct. 2021.

[22] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," *arXiv preprint arXiv:1706.03762*, 2023.

[24] C. D. Kim, B. Kim, H. Lee, and G. Kim, "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.

[25] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 21 450–21 474, 2023.

[26] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[27] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.

[28] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *arXiv preprint arXiv:2306.04186*, 2023.

[29] 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," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[30] W. A. Jassim, J. Skoglund, M. Chinen, and A. Hines, "WARP-Q: Quality prediction for generative neural speech codecs," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 401–405.

[31] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet Audio Distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2019.

[32] Z. Peng, Z. Guo, W. Huang, Y. Wang, L. Xie, J. Jiao, Q. Tian, and Q. Ye, "Conformer: Local features coupling global representations for recognition and detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9454–9468, 2023.