XACLE Challenge 2026: 環境音とテキストにおける主観的意味関連性の自動評価に向けた国際コンペティション *

◎岡本 悠希（東大院），滝沢 力（京産大院），岸 秀（慶大），
金森 勇介（東大院），砺波 紀之（NEC），永瀬 亮太郎（立命館大），
高道 慎之介（慶大/東大），井本 桂右（京都大学）

1 はじめに

テキストや動画など様々な入力から環境音を生成する x-to-audio generation (XTA) が盛んに研究されている [1, 2]。XTA では，生成された環境音が XTA に入力された情報とどの程度意味的に関連しているか（以降，意味関連性）を評価するために，主観および客観評価の両方の側面で評価される [3]。例えば，text-to-audio generation (TTA) [4] に対しては主観評価が多く用いられるが，主観評価はコストが高く，再現性に課題がある。そのため，人間の主観評価と高い相関を持つ自動評価手法の開発が重要である。

本稿では，環境音とテキストにおける主観的意味関連性の自動評価に向けた ICASSP 2026 SP Grand Challenge GC-12 x-to-audio alignment のタスク概要および結果を報告する。Fig. 1 に本コンペティションで扱う課題の概要を示す。本コンペティションでは，環境音と対応するテキストの主観的意味関連性スコアを自動的に予測するモデルを構築し，人間の主観評価と高い相関を持つ自動評価の実現を目指す。このような評価のフレームワークは，人間の指示に忠実に環境音を生成するという XTA の発展において必要不可欠である。さらに，コンピュータを用いて人間同様に環境音とそのテキストの意味的関連性の評価を模擬することは，人間の音に対する知覚を理解する上でも役立つことが期待される。そのため，本タスクは，XTA の研究を促進する上で重要な取り組みである。

2 タスク設定

2.1 XACLE Challenge 2026 dataset

Table 1 に本コンペティションの公式データセットの統計情報を示す。学習，検証，テストデータセットは以下から構成される。

- **環境音とテキストのペアデータ**
  テキストは英語であり，環境音はモノラルで 16-bit，16 kHz の WAV 形式に変換されている。
- **評価者ごとの主観的意味関連性スコア**
  環境音とテキストの主観的意味関連性を 0 (“does not match at all”) から 10 (“matched exactly”) の 11 段階で評価したスコア。各環境音とテキストペアは英語の母語話者によって評価され，学習および検証データセットでは各ペアごとに 4 人，テストデータセットでは 8 人が評価した。
- **各環境音とテキストペアに対する平均スコア**
  各環境音とテキストペアごとの平均スコア。
- **評価者 ID**
  各スコアに対応づいた評価者の ID。

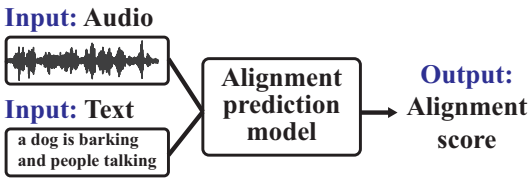


Fig. 1 Overview of task.

本データセットの環境音は自然音と合成音で構成される。学習および検証データセットに使用される自然音は，AudioCaps [5] から選択した。テストデータセットに含まれる自然音は，YouTube から新たに収集したもので，対応するテキストはクラウドソーシングサービスを通じて収集した。合成音は，8 つの TTA モデル（AudioLDM [1]、AudioLDM 2 [4]、Tango [6]、Tango 2 [7]、AudioGen [8]、TangoFlux [9]、Stable Audio [10]、Make-An-Audio 2 [11]）によって生成した。各 TTA モデルへの入力には自然音に対応づいたテキストを使用した。

評価スコア収集の際は，各環境音とテキストのペアを評価者に提示し，“How would you rate the relevance of the audio to the description above?” という質問をして，主観的意味関連性スコアを収集した。なお，コンペティション開催期間中において，テストデータセットでは，環境音とテキストのペア並びに平均スコアのみが参加者に公開され，評価者ごとのスコアと評価者 ID は非公開とした。

2.2 評価指標

参加者から提出されたシステムは，モデルの予測スコアと主観評価スコアの相関係数およびスコアの誤差に基づいて評価される。具体的には，線形相関係数 (LCC)，スピアマンの順位相関係数 (SRCC)，ケンダールの順位相関係数 (KTAU)，平均二乗誤差

---

* XACLE Challenge 2026: An international competition toward automatic evaluation of subjective semantic alignment between environmental sounds and texts by Yuki Okamoto[1], Riki Takizawa[2], Minoru Kishi[3], Yusuke Kanamori[1], Noriyuki Tonami[4], Ryotaro Nagase[5], Shinnosuke Takamichi[3,1], Keisuke Imoto[6] ([1]The University of Tokyo, [2]Kyoto Sangyo University, [3]Keio University, [4]NEC Corporation, [5]Ritsumeikan University, [6]Kyoto University)

Table 1　Statistics of train, validation, and test datasets.

	Training	Validation	Test
#Evaluations	30,000	12,000	12,000
#Audio–text pairs	7,500	3,000	3,000
Audio durations [s]	75,000	30,000	30,000
#total Listeners	2,323	668	1,336

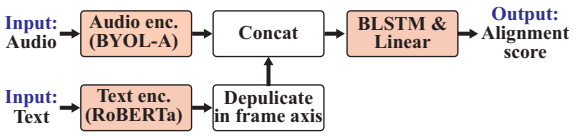


Fig. 2　Architecture of the baseline model.

(MSE) をそれぞれ用いた。$\boldsymbol{y}$ と $\hat{\boldsymbol{y}}$ をそれぞれ主観評価スコアの集合と予測スコアの集合とすると，各評価指標は以下のように計算される。

$$\mathrm{SRCC} = 1 - \frac{6\sum d_i^2}{N(N^2-1)}, \tag{1}$$

$$d_n = \mathrm{rank}(y_n) - \mathrm{rank}(\hat{y}_n). \tag{2}$$

ここで，N はサンプル数，$\mathrm{rank}(\cdot)$ はランクによる並び替えを表す。なお，同じ順位が存在する場合は，平均の順位をそれぞれ割り当てることとする。

$$\mathrm{LCC} = \frac{\sum_{n=1}^{N}(y_n - m_y)(\hat{y}_n - m_{\hat{y}})}{\sqrt{\sum_{n=1}^{N}(y_n - m_y)^2}\sqrt{\sum_{n=1}^{N}(\hat{y}_n - m_{\hat{y}})^2}}. \tag{3}$$

ここで，m_y および $m_{\hat{y}}$ はそれぞれ $\boldsymbol{y}$ および $\hat{\boldsymbol{y}}$ の平均を表す。

$$\mathrm{KTAU} = \frac{N_c - N_d}{\sqrt{(N_c + N_d + N_{tx})(N_c + N_d + N_{ty})}}. \tag{4}$$

ここで，N_c、N_d、N_{tx}，および N_{ty} は，それぞれ予測スコアと主観評価スコアの順位が一致するペアの数，不一致ペアの数，予測スコア内の同順位ペアの数，および主観評価スコア内の同順位ペアの数を表す。

$$\mathrm{MSE} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2, \tag{5}$$

最終順位は SRCC 指標に基づいて決定される。複数のチームが同じ SRCC の値の場合，順位は LCC，KTAU，MSE 指標を用いて決定される。

2.3　タスクのルール

学習用データセットと事前学習済みモデル。参加者は公式の学習データセットに加え，外部データを使用することができる。ただし，その使用を主催者に申告し，許可を得る必要がある。事前学習済みモデルについても同様の規則が適用され，主催者によって承認されたモデルのみが使用可能である。事前学習済みモデルを使用する場合，参加者は当該モデルの学習に使用されたデータについても主催者に申告しなければならない。外部データおよび事前学習済みモデルは，2025 年 5 月 31 日までに公開済みのものに限定される。本コンペティションで使用を許可されたデータセットと事前学習済みモデルのリストは XACLE Challenge の Web ページ[1]において公開されている。

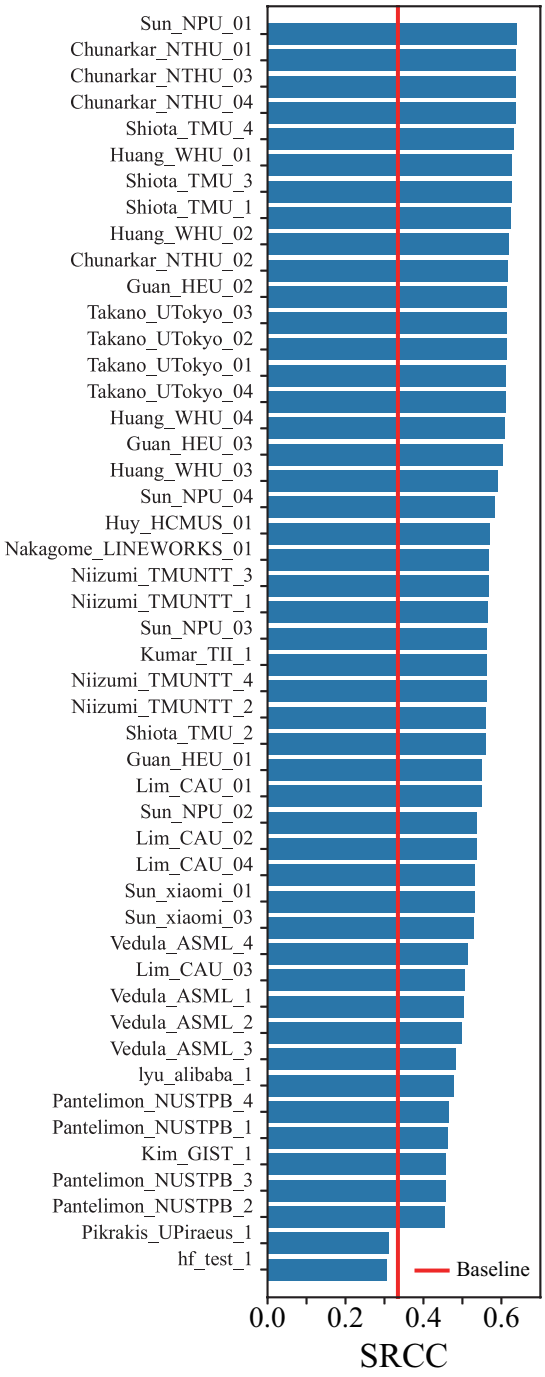


Fig. 3　Overall results.

モデル学習。モデル学習のための新たな環境音，テキスト，スコアの収集は禁止とする。なお，モデルサ

[1] https://xacle.org/description.html

Table 2 Results of top-ranked teams and baseline systems. Team rank indicates the ranking obtained by comparing the best-performing systems from each team, and system rank indicates the ranking obtained by comparing all submitted systems.

Team rank	System rank	Team name	Test				Validation			
			SRCC ↑	KTAU ↑	LCC ↑	MSE ↓	SRCC ↑	KTAU ↑	LCC ↑	MSE ↓
1	1	Sun_NPU	0.6402	0.6873	0.4612	3.0111	0.6680	0.6800	0.4860	3.3340
2	2	Chunarkar_NTHU	0.6382	0.6851	0.4596	2.8256	0.6645	0.6796	0.4829	3.1218
3	5	Shiota_TMU	0.6327	0.6426	0.4564	9.5999	0.6780	0.6716	0.4939	8.3798
4	6	Huang_WHU	0.6264	0.6695	0.4497	2.8369	0.7082	0.7085	0.5208	2.8742
5	11	Guan_HEU	0.6143	0.6770	0.4403	2.8044	0.6455	0.6675	0.4690	3.1925
		Baseline	0.3345	0.3420	0.229	4.8113	0.3844	0.3961	0.2646	4.8361

イズと推論時間に関する制限はない。

2.4 ベースラインモデル

RELATE [12] のモデルを基にベースラインモデルを提供した。Fig. 2 にモデル構造を示す。このモデルは，音響エンコーダ（事前学習済みの BYOL-A [13]），テキストエンコーダ（事前学習済みの RoBERTa [14]），および Long Short-Term Memory (LSTM) ベースのスコア予測器で構成される。環境音とテキストはそれぞれ事前学習済みの音響エンコーダとテキストエンコーダに入力される。各エンコーダで抽出された特徴ベクトルは特徴次元方向に連結された後，スコア予測器へ入力され，スコアが予測される。学習時は，予測スコアと主観評価スコアの clipped MSE 並びに，contrastive loss を損失関数とした。その他のモデルパラメータは RELATE と同様にした。なお，本モデルのソースコードは GitHub にて公開している[2]。

3 結果

3.1 全体の結果

合計 18 チーム，48 システムが提出された。Fig. 3 にテストデータセットに対する各チームの SRCC の結果を示す。18 チーム中 16 チームがベースラインモデルを上回るシステムを構築しており，環境音とテキスト間の主観的意味関連性を予測するタスクにおける著しい進展が示された。

Table 2 に上位 5 チームの各指標における評価結果を示す。特に上位 5 チームは，ベースラインモデルを各指標で大幅に上回るシステムを開発できたことが分かる。

Fig. 4 (a) に全システムでの検証データセットとテストデータセットにおける SRCC の散布図を示す。線形相関を計算すると，0.928 という強い相関が確認され，システム間の相対的な傾向が検証データセットとテストデータセット間で一貫していることが明らかとなった。この結果は，検証データセットとテストデータセット間で大きなドメインシフトが発生していないことを裏付ける結果である。

3.2 提出されたシステムの分析

本コンペティションに提出された 48 システムのうち，34 システムが音響/テキストエンコーダとして CLAP (contrastive language–audio pretraining) [15] 派生のモデルを使用しており，環境音とテキストの共通埋め込みを事前学習するモデルが本タスクにおける主要な基盤となっていることが確認できる。その中でも特に，M2D-CLAP [16] が最も多く 8 チームで使用されていることが確認された。テキストエンコーダのみに着目すると，CLAP 派生のモデルの次に DeBERTa [17] が複数のシステムで使用されていることが確認できた。

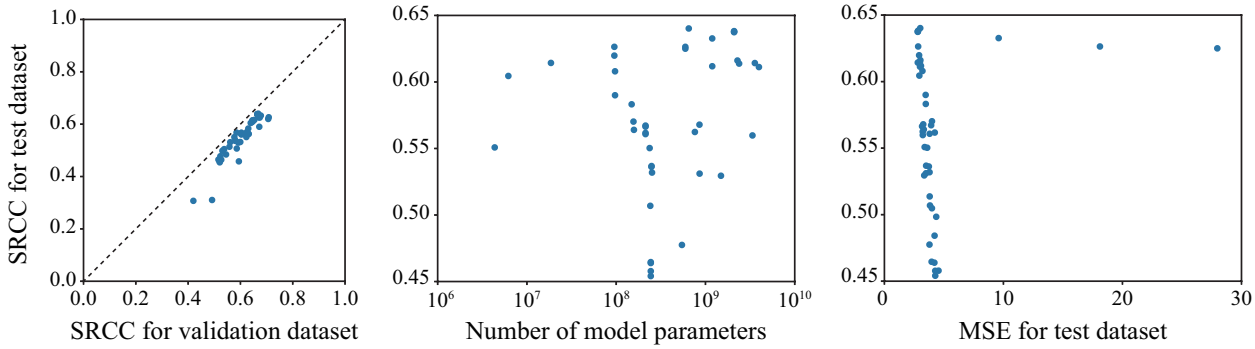
スコア予測器の設計に注目すると，上位のシステムでは Transformer + MLP や Cross-Attention + MLP を用いた構成が多く見られた。また，全体としても Transformer + MLP を採用したシステムが多数を占めていた。これらの結果から，音響エンコーダおよびテキストエンコーダによって特徴抽出した後，時系列的な情報を考慮可能なモデル構造を用いることが有効であった可能性が示唆される。一方で，サポートベクター回帰 (SVR) を使用したチームが 1 チームあり，そのチームが全体 2 位の性能を達成している点も特徴的である。

モデルのパラメータ数と予測性能の関係についても分析した。Fig. 4 (b) にパラメータ数とテストデータセットに対する各システムの SRCC の散布図を示す。線形相関係数を計算すると 0.360 であり，両者の間の相関は弱いことが確認された。SVR を使用したモデルが 2 位を獲得していたこと合わせると，環境音とテキストの主観的意味関連性を予測するために，必ずしも大規模モデルを構築する必要はない可能性が示唆された。一方で，LLM を初めとする Transformer を利用したモデルでは，モデルパラメータのスケーリング則が実験的に示されている [18]。今回のタスクで同様の傾向が確認されなかった理由として，学習のためのデータ数の少なさが起因している可能性が考えられる。

損失関数の傾向としては，MSE を採用したシステムが最も多く確認された。一方で，順位情報を考慮した損失関数を組み込んだシステムも複数存在し，それらは上位のシステムに多い傾向が確認された。全体としては，MSE を基本としつつ，他の損失関数を組み合わせる設計が多く確認された。

外部データセットの使用に関しては，48 システム中 10 システムが外部データセットを使用していた。外部データを使用する場合には，AudioCaps や

[2] https://github.com/XACLE-Challenge/the_first_XACLE_challenge_baseline_model



(a) SRCC on validation vs. test datasets　(b) Relationship between number of parameters and SRCC　(c) Relationship between SRCC and MSE

Fig. 4　Results of the relationships among evaluation metrics

CLOTHO [19] など，環境音とテキストが対応づいたデータセットをモデルの事前学習に利用するケース多く見られた。

テストデータセットに対する SRCC と MSE の関係性についても分析した。Fig. 4 (c) にシステムごとの SRCC と MSE の散布図に示す。図を確認すると，SRCC が高いにも関わらず MSE の値が極端に大きいシステムが 3 つ確認できる。これらのシステムでは，ランキング学習のみが採用されているおり，予測値と主観評価スコア間の絶対的な誤差が考慮されていないため，MSE の値が大きくなったと考えられる。一方で，MSE が極端に大きい 3 つのシステムを除外して 2 つの指標間で SRCC を計算すると −0.884 となり，両者の間には強い負の相関関係が確認された。

4　コンペティション全体を通して

今回，参加者が検証データに対する評価結果と自身の順位をリアルタイムで確認できるように，Kaggle[3] のプラットフォームも利用した。多くの参加者がプラットフォームを使用した一方で，本来 Kaggle で行われる参加者同士の議論は見られなかった。次回以降開催する際には，コンペティションのための slack ワークスペースの導入など，参加者同士が活発に交流できる仕組みづくりも考えたい。

5　まとめ

本稿では, 環境音とテキストの主観的意味関連性の自動評価に向けた国際コンペティションである ICASSP 2026 SP Grand Challenge GC-12 の概要並びに結果を報告した。提出されたほとんどのシステムがベースラインを上回る結果となり，上位システムでは SRCC で約 0.3 の改善を達成した。提出されたシステム全体を通して，大規模事前学習済みモデルを音響/テキストエンコーダとして使用されていた。一方で，スコア予測器において SVR を使用したシステムが 2 位を獲得するなど，現タスクにおいては必ずしも大規模モデルの構築が必要ではないことが示唆された。本コンペティションの結果および参加者から提出されたテクニカルレポートが，今後の本タスクの発展に寄与することを願う。

謝辞　本研究の一部は，JSPS 科研費 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。また，合成音の収集に尽力頂いた苗村公明氏に深謝申し上げます。

参考文献

[1] H. Liu, *et al.*, Proc. ICML, pp. 21450–21474, 2023.
[2] L. Ruan, *et al.*, Proc. CVPR, pp. 10219–10228, 2023.
[3] K. Choi, *et al.*, Proc. DCASE, pp. 16–20, 2023.
[4] H. Liu, *et al.*, IEEE/ACM TASLP, vol. 32, pp. 2871–2883, 2024.
[5] C. D. Kim, *et al.*, Proc. NACCL, pp. 119–132, 2019.
[6] D. Ghosal, *et al.*, Proc. ACM MM, pp. 3590–3598, 2023.
[7] N. Majumder, *et al.*, Proc. ACM MM, pp. 564–572, 2024.
[8] F. Kreuk, *et al.*, Proc. ICLR, 2023.
[9] C.-Y. Hung, *et al.*, arXiv preprint arXiv:2412.21037, 2024.
[10] https://stableaudio.com/
[11] J. Huang, *et al.*, arXiv preprint arXiv:2305.18474, 2023.
[12] Y. Kanamori, *et al.*, Proc. INTERSPEECH, pp. 3155–3159, 2025.
[13] D. Niizumi, *et al.*, IEEE/ACM TASLP, vol. 31, pp. 137–151, 2023.
[14] Y. Liu, *et al.*, arXiv preprint, arXiv:1907.11692, 2019.
[15] Y. Wu, *et al.*, Proc.ICASSP, pp. 1–5, 2023.
[16] D. Niizumi, *et al.*, Proc. INTERSPEECH, pp. 57–61, 2024.
[17] P. He, *et al.*, Proc. ICLR, 2021.
[18] J. Kaplan, *et al.*, arXiv preprint arXiv:2001.08361, 2020.
[19] K. Drossos, *et al.*, Proc. ICASSP, pp. 736–740, 2020.

[3] https://www.kaggle.com/