TRUST YOUR CLAP: TMU-NTT XACLE CHALLENGE 2026 TECHNICAL REPORT

Daisuke Niizumi[†‡], *Daiki Takeuchi*[†], *Masahiro Yasuda*[†], *Binh Thien Nguyen*[†], *Noboru Harada*[†], *and Nobutaka Ono*[‡]

[†]NTT Inc., Japan, [‡]Tokyo Metropolitan University, Japan

ABSTRACT

This report presents our solution for the XACLE Challenge 2026. The challenge aims to address the low correlation between CLAPScore and human evaluation. However, this problem setting implies that while high-scoring audio-caption pairs form a meaningful joint probability distribution, low-scoring pairs do not. This highlights a fundamental difficulty in building a machine learning system that can model a subjective human being. We believe that CLAP remains useful for addressing the issue, as it is trained to directly evaluate such joint probabilities. Our solution used a frozen M2D-CLAP model to predict human scores as a baseline, and employed CLAP-based score estimation when modeling human judgment was difficult. The validation results demonstrate that CLAP remains trustworthy.

Index Terms— XACLE, Trustworthy CLAP

1. INTRODUCTION

Previous studies, such as RELATE [1] and Human-CLAP [2], have highlighted the problem of discrepancies between the CLAPScore [3] and human evaluations. The XACLE Challenge 2026 [4][1] requires systems that address this issue.

We consider that the essence of the problem lies in the fact that, when human evaluation scores are low, there exist numerous audio–text pairs of this type, making the prediction task ill-posed. Rather, we hypothesize that treating such cases as anomalies and approaching the problem as one of anomaly detection may lead to a solution. Accordingly, we utilize the CLAPScore, which can directly assess the similarity of a pair, to detect whether a pair is anomalous. Our approach demonstrates that leveraging the CLAPScore improves performance, and, somewhat ironically, indicates that the CLAPScore remains sufficiently reliable for this purpose.

2. OUR SYSTEM

Our system, shown in Fig.1, leverages the CLAP representations provided by M2D-CLAP [5], and combines two complementary approaches: (1) score prediction through regression, and (2) score estimation using CLAPScore. Each subsystem utilizes the CLAP features extracted from the input audio x_a and text x_t, denoted as z_a and z_t, respectively. These features (768-d) serve as the primary inputs to the subsystems.

In regression-based score prediction, the two feature vectors z_a and z_t are used to directly predict the human evaluation score y_{GT}. Since we keep the M2D-CLAP parameters frozen, we employ a regression network with moderate model capacity. Specifically, we use a two-layer standard Transformer encoder that processes the concatenated input sequence $[z_t; z_a]$. The 768-d output corresponding to z_t is then fed into an MLP with a single hidden layer of 512 units to produce the predicted scalar score $\hat{y}_r$.

In CLAPScore-based score estimation, the CLAPScore C is calculated as the cosine similarity between z_a and z_t. The CLAP-based estimated score $\hat{y}_c$ is then derived as

$$\hat{y}_c = 10\,\text{min-max}(C), \tag{1}$$

where the min–max normalization ensures that the similarity values are mapped onto a consistent 0–10 scale.

The final predicted score $\hat{y}$ is determined based on the value of C relative to its distribution. For each sample i, $\hat{y}^{(i)}$ is defined as follows:

$$\hat{y}^{(i)} = \begin{cases} \hat{y}_r^{(i)}, & \text{if } C^{(i)} > \mu - 2\sigma, \\ \hat{y}_c^{(i)}, & \text{if } C^{(i)} \leq \mu - 2\sigma, \end{cases} \tag{2}$$

where μ and σ denote the mean and standard deviation of the CLAPScore C, respectively. The threshold $\mu - 2\sigma$ was determined empirically. In this way, the CLAPScore is utilized as a confidence measure for score prediction: when the confidence is low (i.e., $C \leq \mu - 2\sigma$), the CLAP-based estimation is used as the final score.

For the training of the regression-based score prediction, we employed an approach to improve prediction accuracy for samples with higher human evaluation scores. The mean squared error (MSE) was used as the loss function, and a per-sample loss weight w_{loss} was assigned to implement the approach. Four types of loss weights were prepared, and the weight for each sample i, denoted $w_{\text{loss}}^{(i)}$, was set based on the human score $y_{GT}^{(i)}$ as illustrated in Fig.2. Type0 applies no bias. Type1 is designed to apply moderate weighting to higher

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

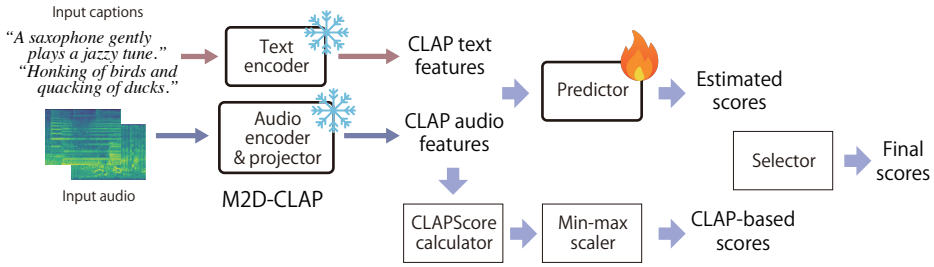

Fig. 1. System diagram for our solution.

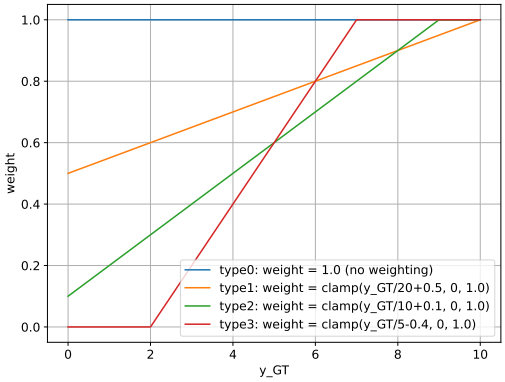

Fig. 2. Loss weight type comparison.

scores, whereas type2 and type3 place greater emphasis on higher scores.

For data augmentation, we employed SpecAugment [6], which masks parts of the input spectrogram. In addition, we used a cosine annealing learning rate scheduler and the LARS optimizer implemented based on the EVAR[2] codebase. For subsequent testing, we selected the checkpoint that achieved the highest Spearman's rank correlation coefficient (SRCC) between $\hat{y}$ and y_{GT} on the validation set.

3. RESULTS

We conducted experiments on the validation set, averaging the results over five runs. The results show that while regression alone achieved an SRCC of 0.5907, incorporating CLAP-based score estimation improved performance in all cases, yielding SRCCs of 0.5996, 0.6006, 0.5994, and 0.5919 for loss weight types 0–3. This demonstrates that combining CLAP-based similarity with regression can provide an improvement of approximately 0.01. Among the loss weights, type1, which applies moderate weighting to higher scores, performed the best.

[2]https://github.com/nttcslab/eval-audio-repr

4. CONCLUSION

This report shows that, contrary to previous assumptions, CLAP can still be effectively used to address discrepancies between the CLAPScore and human evaluations. It also suggests a broader potential for contributing to the solution of this and other related problems by enabling CLAP to more rigorously assess audio-text pairs.

5. ACKNOWLEDGMENT

This work was partially supported by JST Strategic International Collaborative Research Program (SICORP), Grant Number JPMJSC2306, Japan.

6. REFERENCES

[1] Yusuke Kanamori, Yuki Okamoto, Taisei Takano, Shinnosuke Takamichi, and Hiroshi Saruwatari Yuki Saito, "RELATE: Subjective evaluation dataset for automatic evaluation of relevance between text and audio," in *Proc. Interspeech*, 2025, pp. 3155–3159.

[2] Taisei Takano, Yuki Okamoto, Yusuke Kanamori, Yuki Saito, Ryotaro Nagase, and Hiroshi Saruwatari, "Human-CLAP: Human-perception-based contrastive language-audio pretraining," in *Proc. APSIPA ASC*, 2025.

[3] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao, "Make-an-audio: text-to-audio generation with prompt-enhanced diffusion models," in *Proc. ICML*, 2023.

[4] Yuki Okamoto, Riki Takizawa, Minoru Kishi, Yusuke Kanamori, Noriyuki Tonami, Ryotaro Nagase, Shinnosuke Takamichi, and Keisuke Imoto, "XACLE Challenge 2026: The first x-to-audio alignment challenge," 2025.

[5] Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada, "M2D-CLAP: Exploring general-purpose audio-language representations beyond clap," *IEEE Access*, vol. 13, pp. 163313–163330, 2025.

[6] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019, pp. 2613–2617.