D2-0830_L1.11

Human-CLAP: Human-perception-based contrastive language–audio pretraining

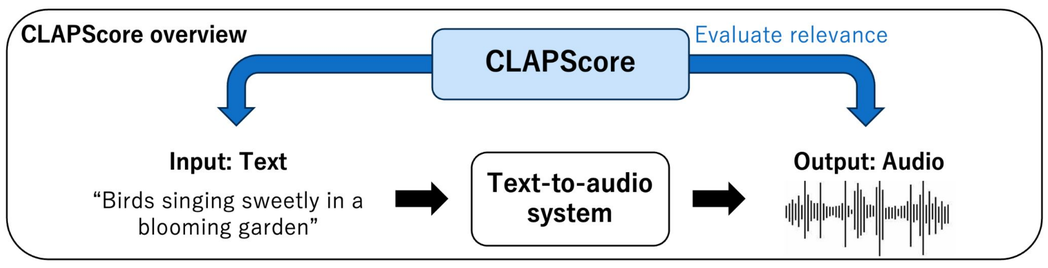
Our project

Taisei Takano*, Yuki Okamoto*, Yusuke Kanamori*, Yuki Saito*, Ryotaro Nagase†, Hiroshi Saruwatari*
*The University of Tokyo, Japan, †Ritsumeikan University, Japan

① Background

Evaluating text–audio semantic relevance

■ An important aspect when evaluating text-to-audio (TTA)
- Subjective evaluation: Human-scored similarity
  - Extremely costly in time and money
- Objective evaluation: CLAPScore



■ Issue
- Relationship with human perception is unclear

→ How reliable is CLAPScore?

Purpose and contribution

Analyzed CLAPScore
- **Correlation between human-scored similarity and CLAPScore**

→ **Low correlation**

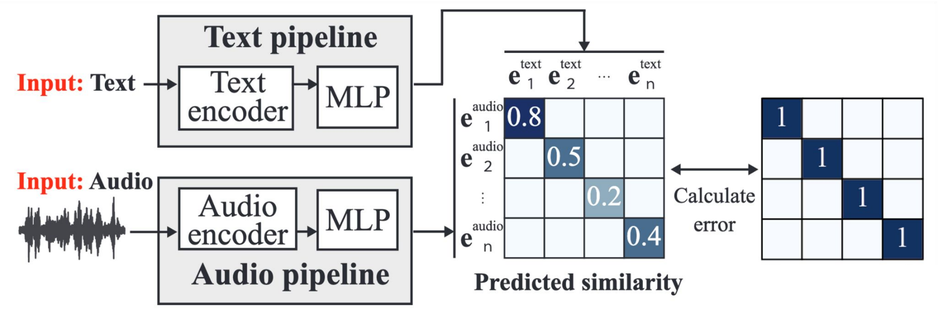
The correlation improvement with human-scored similarity
- **Fine-tuned CLAP using a small amount of human-scored text–audio similarity**

→ **Effectively improved the correlation, enabling CLAPScore to align more closely with human perception**

② CLAPScore

Evaluate text–audio similarity using CLAP

■ Contrastive language–audio pretraining (CLAP) [1]
- **Trained to bring paired text–audio embeddings closer together**



■ CLAPScore [2]
- **Calculate the cosine similarity between text and audio embeddings obtained from CLAP**

$$\mathrm{CLAPScore} = \max\left(\frac{\mathbf{e}^{\mathrm{audio}} \cdot \mathbf{e}^{\mathrm{text}}}{\|\mathbf{e}^{\mathrm{audio}}\|\|\mathbf{e}^{\mathrm{text}}\|}, 0\right), \quad \mathbf{e}^{\mathrm{audio}}: \text{Audio embedding}, \quad \mathbf{e}^{\mathrm{text}}: \text{Text embedding}$$

Analyzed CLAPScore

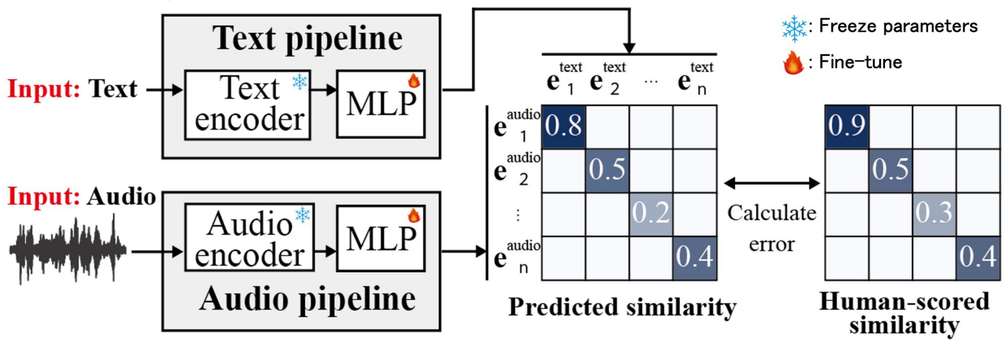
■ Correlation between CLAPScore and human-scored similarity
- Metric: Spearman's rank correlation coefficient (SRCC)
- Evaluated LAION CLAP [3] on RELATE [4] test set

SRCC: 0.280
Insufficient correlation with human evaluations

③ Proposed Method: Human-CLAP

CLAP model based on human perception

■ Fine-tuned CLAP **with human-scored similarity**
- Minimize the difference between predicted score and human-scored similarity



Combining regression and contrastive learning

■ Regression loss
- Mean absolute error (MAE)

$$L_{MAE} = \frac{1}{N}\sum_{i=1}^{N}|a_i - y_i|$$

a_i: Human-scored similarity, y_i: Predicted similarity, N: Batch size

■ Contrastive learning
- Weight symmetric cross entropy loss (wSCE) (Proposed)
- Symmetrical InfoNCE [5] **weighted by human-scored similarity**

$$L_{wSCE} = -\frac{1}{2N}\sum_{i=1}^{N} a_i \left(\log\left(\frac{\exp(e_i^{\mathrm{text}} \cdot e_i^{\mathrm{audio}} / \tau)}{\sum_{j=1}^{N}\exp(e_i^{\mathrm{text}} \cdot e_j^{\mathrm{audio}} / \tau)}\right) + \log\left(\frac{\exp(e_i^{\mathrm{audio}} \cdot e_i^{\mathrm{text}} / \tau)}{\sum_{j=1}^{N}\exp(e_i^{\mathrm{audio}} \cdot e_j^{\mathrm{text}} / \tau)}\right)\right)$$

a_i: Human-scored similarity, N: Batch size, τ: Temperature, $\mathbf{e}_i$: Embedding

④ Evaluation

Experimental setup

■ Dataset: RELATE [4]
- 11-point scale **human-scored similarity of text–audio pairs**
  - 0 (low similarity) ~ 10 (high similarity)

  → Rescaled to the range of 0 to 1 to use as the target score
- **Natural and synthesized audio samples** included
  - Natural: AudioCaps [6]
  - Synthesized: AudioLDM [7], AudioLDM2 [8], Tango [9], Tango2 [10]
- Each pair evaluated by an average of four listeners

Train: 1,880 pairs, Validation: 512 pairs,
Test: 2,730 pairs

■ Pretrained encoders from LAION CLAP [3]
- Text encoder: RoBERTa [11], Audio encoder: HTS-AT [12]

Evaluation metrics

■ Correlation between CLAPScore and human-scored similarity
- Spearman's rank correlation coefficient (**SRCC**)
- Linear correlation coefficient (LCC)
- Kendal's rank correlation coefficient (KTAU)

■ Score difference : Mean squared error (MSE)

Overall results

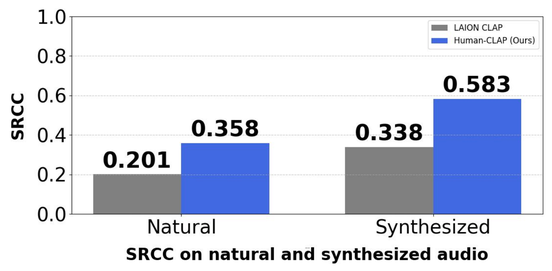
Human-CLAP improved the correlation between CLAPScore and human-scored similarity

Model	SRCC ↑	LCC ↑	KTAU ↑	MSE ↓
Human-CLAP (ours)				
wSCE + MAE	**0.457**	**0.481**	**0.320**	0.057
wSCE	0.383	0.410	0.265	0.063
MAE	0.453	0.472	0.317	**0.051**
Baseline				
LAION CLAP	0.280	0.294	0.192	0.068
MS CLAP	0.278	0.296	0.192	0.078

SRCC: +0.177

Results on natural and synthesized audio

Correlation improved for both natural and synthesized audio



Conclusion

Objective evaluation metric for text–audio similarity

■ Conventional CLAPScore had a **low correlation** with human-scored similarity

■ **Human-CLAP** effectively **improved the correlation**, enabling CLAPScore to align **more closely with human perception**

Future work

■ Analyze the prediction tendency of CLAPScore
- Depending on the types of sound events

■ Transfer Human-CLAP to other text–audio tasks **related to human perception** (e.g. TTA)

References
[1] B. Elizalde, et al., 2023. [2] R. Huang, et al., 2023. [3] Y. Wu, et al., 2023. [4] Y. Kanamori, et al., 2025. [5] A. Oord, et al., 2018. [6] C. D. Kim, et al., 2019. [7] H. Liu, et al., 2023. [8] H. Liu, et al., 2024. [9] D. Ghosal, et al., 2023. [10] N. Majumder, et al., 2024. [11] Y. Liu, et al., 2019. [12] K. Chen, et al., 2022.

Acknowledgements
The work was supported by JSPS KAKENHI Grant Number 24K23880, 25K21221, ROIS NII Open Collaborative Research 2025-(251S4-22735), JST Moonshot Grant Number JPMJMS2237.