XACLE CHALLENGE 2026: THE FIRST X-TO-AUDIO ALIGNMENT CHALLENGE

Yuki Okamoto[1]*, Riki Takizawa*[2]*, Minoru Kishi*[3]*,Yusuke Kanamori*[1]*,*
Noriyuki Tonami[4]*, Ryotaro Nagase*[5]*, Shinnosuke Takamichi*[3,1]*, and Keisuke Imoto*[6]

[1]The University of Tokyo, Japan, [2]Kyoto Sangyo University, [3]Keio University,Japan,
[4]NEC Corporation, Japan, [5]Ritsumeikan University, Japan, [6]Kyoto University, Japan

ABSTRACT

We present the task description of the ICASSP 2026 Grand Challenge GC-12: "x-to-audio alignment." The scope of this challenge is to predict the semantic alignment of a given general audio and text pair. In this challenge, our goal is to build a model that automatically predicts the semantic alignment from the pair for evaluating text-to-audio generation (TTA). The aim is to develop a method for automatic evaluation that correlates highly with human subjective evaluations. The challenge results will be added after the submission deadline.

***Index Terms*—** x-to-audio generation, text-to-audio generation, objective evaluation, subjective evaluation, audio–language model

1. INTRODUCTION

The generation of general audio from various inputs, such as text and video (x-to-audio generation), has been actively studied [1]. In x-to-audio generation, both subjective and objective evaluations of how well the output matches the input are extremely important. For instance, in the evaluation of text-to-audio generation (TTA), methods have been proposed to evaluate the alignment between audio and text objectively. However, it has been pointed out that these methods often have a low correlation with human subjective evaluations [2].

Figure 1 illustrates the overview of the task we present. In this task, we aim to build a model that automatically predicts the alignment score between audio and text for text-to-audio evaluation, specifically, to achieve objective evaluations that correlate highly with human subjective assessments. Our ultimate goal is to faithfully generate audio from human instructions, and evaluating input–output alignment is crucial for this advancement. Moreover, the development of automated evaluation methods that are strongly correlated with human evaluations is helpful for understanding human audio perception. Furthermore, in recent years, noisy data from the internet have often been used for training models such as text-to-audio generation. This task is also expected to be useful for screening such noisy data. Thus, this task is a very important initiative for advancing tasks that deal with text and audio.

This work was supported by JSPS KAKENHI Grant Number 24K23880, 25K21221, Support Center for Advanced Telecommunications Technology Research Foundation, and JST Moonshot Grant Number JPMJMS2011. The authors also thank Koumei Naemura for his support in collecting audio samples.

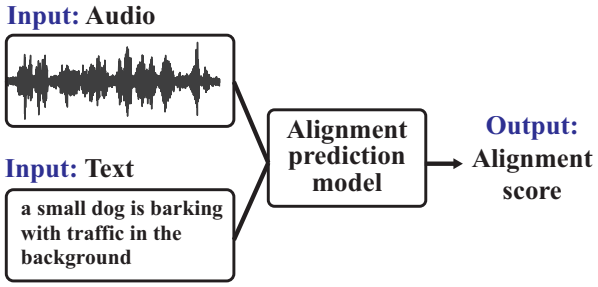

Fig. 1. Overview of tasks

2. TASK SETUP

2.1. Dataset

2.1.1. Training and validation data

The training and validation data consist of the following:

- **Audio–text pairs**
  Each text is written in English, and all audio samples were converted to mono 16-bit 16 kHz RIFF WAV format.
- **11-point semantic-alignment scores between audio and text per listener**
  This dataset includes subjective evaluation scores for semantic alignment between audio and text. The semantic-alignment score is on an 11-point scale from 0 ("does not match at all") to 10 ("matched exactly"). Each audio–text pair was evaluated by four native English speakers.
- **Average semantic-alignment scores per audio–text pair (average scores of each audio–text pair)**
  This dataset includes average semantic-alignment scores for each audio–text pair.
- **Listener IDs who gave the semantic-alignment scores**

The details of the training and validation data will be released after the challenge ends.

2.1.2. Test data

The test data consist of **3,000 audio–text pairs**, and each audio–text pair was evaluated by eight listeners. The listeners for the test data were different from those who evaluated the training and validation data. The details of the test data will be released after the conclusion of the challenge.

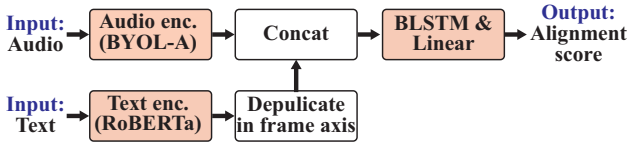

Fig. 2. Architecture of baseline model

2.2. Evaluation metrics

The purpose of this challenge is to develop a method for automatic evaluation that correlates highly with human subjective evaluations. Therefore, we will use metrics that demonstrate correlations and differences from human evaluation scores. Submissions will be evaluated on the basis of the correlation coefficient and score error between predicted and average-semantic-alignment scores. Specifically, the metrics include the linear correlation coefficient (LCC), Spearman's rank correlation coefficient (SRCC), Kendall's rank correlation coefficient (KTAU), and mean squared error (MSE) referring to the VoiceMOS Challenge [3]. When $\boldsymbol{y}$ represents the average-semantic-alignment scores and $\hat{\boldsymbol{y}}$ represents the predicted scores for each audio–text pair, the evaluation metrics are calculated as follows.

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

$$d_i = \mathrm{rank}(y_i) - \mathrm{rank}(\hat{y}_i) \tag{2}$$

where n and $\mathrm{rank}(\cdot)$ denote the number of samples and sorting by rank. If there are ties, the average rank is assigned to each of the tied values.

$$\mathsf{LCC} = \frac{\sum_{n=1}^{N}(y_i - m_y)(\hat{y}_i - m_{\hat{y}})}{\sqrt{\sum_{n=1}^{N}(y_i - m_y)^2}\sqrt{\sum_{n=1}^{N}(\hat{y}_i - m_{\hat{y}})^2}}, \tag{3}$$

where m_y, $m_{\hat{y}}$, and N denote the mean of the vector $\boldsymbol{y}$, $\hat{\boldsymbol{y}}$, and number of samples, respectively.

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

where N_c, N_d, N_{tx} ,and N_{ty} denote the number of concordant pairs where the ranks of the predicted scores and average semantic-alignment scores, discordant pairs, tied pairs for the x-axis variable, and tied pairs for the y-axis variable.

$$\mathsf{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{5}$$

The final ranking is determined based on the SRCC metric. If multiple teams are tied, the standings will be determined using the LCC, KTAU, and MSE metrics.

2.3. Task rules

Use of dataset and pre-trained model. Participants can use external data in addition to the training set provided by the organizers, if they declare its use to the organizers and receive permission. Pre-trained models are subject to the same rule: only the models declared to the organizers and approved by them may be used. If participants use pre-trained models, they must also declare to the organizers what data were used to train those models.

Table 1. Results of baseline model for validation data

	SRCC ↑	KTAU ↑	LCC ↑	MSE ↓
Baseline	0.384	0.396	0.264	4.836

External data and pre-trained models are limited to those that are open-access. External data are limited to openly available datasets that have already been published. It is prohibited to collect new audio samples, texts, and scores for use in model training.

Model training. To ensure experimental reproducibility, ensembling results from multiple models is prohibited. However, internal ensembling within a single model, such as a mixture of experts, is permitted. There will be no specific restrictions on the model size and inference time for the prediction models participants create.

2.4. Baseline system and results

We have provided a supervised score prediction model, similar to the baseline model of the RELATE [4]. Figure 2 shows the model architecture. The baseline model consists of audio and text encoders, and an long sohrt-term memory (LSTM)-based score predictor. We used pre-trained BYOL-A and RoBERTa for audio and text encoders, respectively. The audio x and text l are input to the pre-trained audio and text encoders, respectively. Since this is the first challenge focusing on audio–text alignment, we adopted a simple baseline model. The source code of the baseline model is publicly available[1].

Table 1 shows the evaluation results for validation data by each metric. The results on the test data will be released after the challenge ends.

3. CHALLENGE RESULTS

The results of each team on the validation and test data will be made available after the end of the challenge.

4. REFERENCES

[1] H. Liu et al., "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[2] T. Takano et al., "Human-clap: Human-perception-based contrastive language-audio pretraining," in *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2025, pp. 131–136.

[3] W.-C. Huang et al., "The voicemos challenge 2024: Beyond speech quality prediction," in *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 803–810.

[4] Y. Kanamori et al., "RELATE: Subjective evaluation dataset for automatic evaluation of relevance between text and audio ," in *Proc. Interspeech*, 2025, pp. 3155–3159.

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