AudioBERTScore: Objective Evaluation of Environmental Sound Synthesis Based on Similarity of Audio Embedding Sequences

Minoru Kishi MINORUKISHI1091@KEIO.JP
Keio University, Japan

Ryosuke Sakai
Keio University, Japan

Shinnosuke Takamichi SHINNOSUKE_TAKAMICHI@KEIO.JP
Keio University, Japan
The University of Tokyo, Japan

Yusuke Kanamori
The University of Tokyo, Japan

Yuki Okamoto
The University of Tokyo, Japan

Editors: Tatsuya Komatsu, Keisuke Imoto, Xiaoxue Gao, Nobutaka Ono, Nancy F. Chen

Abstract

We propose a novel objective evaluation metric for synthesized audio in text-to-audio generation (TTA), aiming to improve the performance of TTA models. In TTA, subjective evaluation of synthesized sounds is important; however, conducting it requires significant monetary and time costs. Therefore, objective evaluation such as mel-cepstral distortion are used, but the correlation between these objective metrics and subjective evaluation values is weak. Our proposed objective evaluation metric, AudioBERTScore, calculates the similarity between embedding of the synthesized and reference sounds. The method is based not only on the max-norm used in conventional BERTScore but also on the p-norm to reflect the non-local nature of environmental sounds. Experimental results show that scores obtained by the proposed method have a higher correlation with subjective evaluation values than conventional metrics.

Keywords: text-to-audio, evaluation metric, semantic similarity, AudioBERTScore

1. Introduction

Text-to-audio generation (TTA) models are deep learning models that synthesize environmental sounds from text inputs such as “a small dog is barking.” Synthesized audio is used for media content creation Marrinan et al. (2024) and for expressing characters’ emotions. The performance of TTA models is primarily evaluated based on the synthesized audio, and subjective evaluation is considered the most important Okamoto et al. (2022). In fact, the final evaluation of TTA models in the DCASE 2024 Challenge Task7[1] is based on subjective evaluation.

Subjective evaluation of synthesized audio is typically conducted in terms of overall quality (OVL) Hansen and Pellom (1998) and relevance to the input text (REL) Okamoto

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

et al. (2022), but it requires considerable time and financial costs. Therefore, objective metrics such as CLAPScore Xiao et al. (2024) and mel-cepstral distortion (MCD) Kubichek (1993) have been proposed. It is essential to examine their correlation with subjective evaluations Okamoto et al. (2022); however, CLAPScore shows a low correlation with subjective evaluation scores Takano et al. (2025), and the correlation between MCD and subjective evaluation scores has not been examined in the context of TTA.

In this study, we propose a new objective evaluation metric, AudioBERTScore, as shown in Figure 1. This metric uses both synthesized and reference audio. Embedding sequences are extracted from each audio using an audio foundation model such as ATST-Frame Li et al. (2024), and the evaluation score is estimated based on the similarity between the two sequences. Experimental results show that among objective metrics using both synthesized and reference audio, the score of the proposed metric correlates most strongly with subjective evaluation scores. The code will be made publicly available on the project page[2].

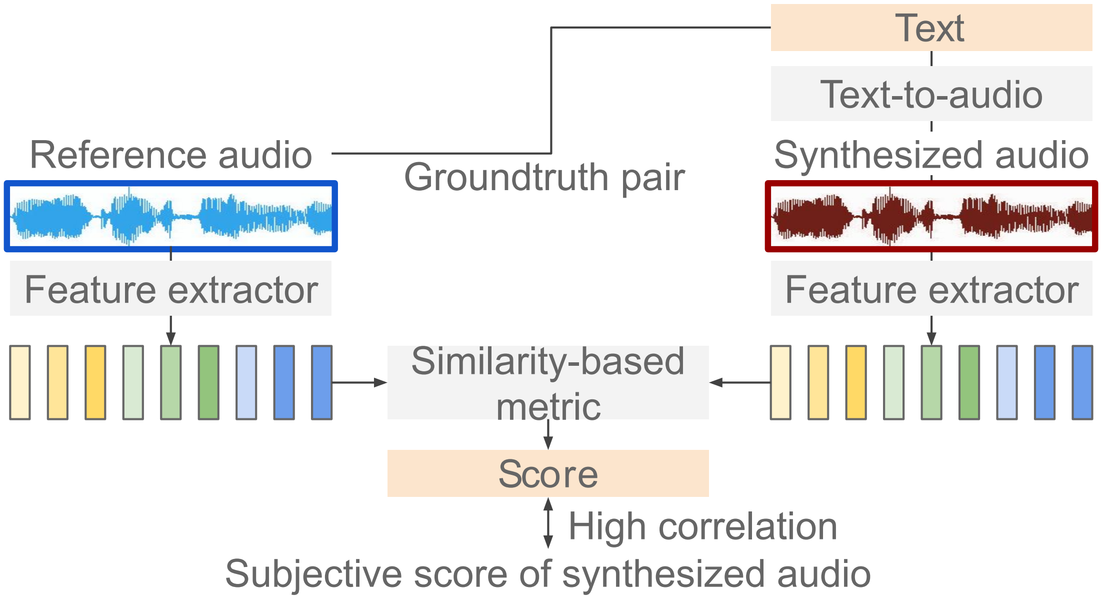


Figure 1: Overview of the proposed AudioBERTScore.

2. Related work

2.1. Evaluation metrics for synthesized audio of TTA

Subjective evaluations include OVL and REL Choi et al. (2023). The former evaluates the perceptual quality of the synthesized audio, including its naturalness, while the latter assesses the correspondence between the input text and the synthesized audio. Subjective evaluation is a useful means of quantifying perceived quality but requires financial costs.

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

Therefore, objective evaluation methods that can predict scores correlated with subjective scores have been studied. In this paper, we classify the methods according to the following criteria:

- **Supervised training:** A direct approach is to train a model to predict scores from the synthesized audio, with supervised training on text, synthesized audio, and subjective scores. We distinguish methods by the existence of such data and training.
- **Reference:** Whether reference data is used in score estimation, and if so, what type, e.g., text-reference-aware, audio-reference-aware, or reference-free.
- **Pretrained model:** Whether a pretrained model is used in the estimation. If pretrained models are used, methods are further categorized by the type of data used in pretraining.

We classify existing methods on the basis of these criteria. They are shown in Table 1 and as follows.

Metric	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)

Table 1: Classification of objective evaluation metrics for synthesized audio

- **MCD, WARP-Q:** MCD Berndt and Clifford (1994) and WARP-Q Han and Zhang (2022) compute the distance between the signal processing features of synthesized and reference audio. Since these scores can be calculated solely from the reference audio, they have the advantage of language independency of the input text. However, the correlation between them and subjective scores has not yet been investigated.
- **FAD:** This method Kilgour et al. (2019) measures the Fréchet distance Fréchet (1906) between the distributions of embedding features extracted from reference and synthesized audio using a pretrained audio classification model. Unlike feature-based distances such as MCD or WARP-Q, FAD captures higher-level perceptual differences in audio quality and naturalness. However, it requires a sufficiently large set of samples to yield stable estimates Kilgour et al. (2019).
- **RELATE benchmark:** This method Kanamori et al. (2025) was proposed as a supervised training model trained on sets of synthesized or natural audio, text, and REL scores. Supervised training can achieve high correlation but requires large-scale datasets of subjective scores.

- **PAM:** This method Deshmukh et al. (2024) uses contrastive learning of text and audio to estimate quality scores from text prompts referring to audio quality and the synthesized audio. PAM does not require reference, but since it depends on pretraining with paired text-audio data, it can only be used for languages where such data exists.
- **CLAPScore:** This method Xiao et al. (2024) uses a pretrained contrastive language—audio pretraining (CLAP) model (e.g., MS-CLAP Elizalde et al. (2023), LAION-CLAP Wu et al. (2023)) to score the similarity between the textual prompt (input transcript) and the synthesized audio. Like PAM, it does not require a reference; however, because it relies on pretraining with paired text—audio data, its applicability depends on the language coverage of the pretraining corpus.

Our AudioBERTScore uses reference audio without supervised training, like MCD and WARP-Q. While maintaining the strength of language independence, it aims for high correlation with subjective scores by improving feature extraction and similarity calculation using foundation models.

2.2. BERTScore

BERTScore Zhang et al. (2020) in natural language processing calculates a series of contextual embedding vectors from both the generated and reference sentences and evaluates similarity between these series. It achieves high correlation with subjective evaluations by using the bidirectional encoder representations from Transformers (BERT) Devlin et al. (2019) self-supervised learning model for embedding, and by computing similarity with forced alignment of the sequences.

2.3. SpeechBERTScore

SpeechBERTScore Saeki et al. (2024) in speech processing, which is originated from BERTScore Zhang et al. (2020), was proposed to automatically evaluate synthesized speech in text-to-speech. SpeechBERTScore successfully applies this framework to synthesized speech by replacing BERT with speech-specific foundation models Hsu et al. (2021).

Our AudioBERTScore also follows this trend. Specifically, it uses audio foundation models for the feature extraction. Furthermore, we design similarity calculation for environmental sound.

3. Proposed objective evaluation metric

3.1. Feature extraction and similarity matrix

We first obtain the embedding sequences for both the synthesized and reference audio. Let the waveform of the synthesized audio be represented as $\boldsymbol{s} = (s_t \in \mathbb{R} \mid t = 1, \ldots, T_{\mathrm{syn}})$, and that of the reference audio as $\boldsymbol{r} = (r_t \in \mathbb{R} \mid t = 1, \ldots, T_{\mathrm{ref}})$. $T_{\mathrm{syn}} \neq T_{\mathrm{ref}}$ in general.

The embedding sequences extracted from $\boldsymbol{s}$ and $\boldsymbol{r}$ using a feature extractor are represented as:

$$\tilde{S} = (\tilde{\boldsymbol{s}}_n \in \mathbb{R}^D \mid n = 1, \ldots, L_{\mathrm{syn}}),$$
$$\tilde{R} = (\tilde{\boldsymbol{r}}_n \in \mathbb{R}^D \mid n = 1, \ldots, L_{\mathrm{ref}}).$$

These are computed as $\tilde{S} = \mathrm{Encoder}(\boldsymbol{s};\theta)$, $\tilde{R} = \mathrm{Encoder}(\boldsymbol{r};\theta)$. θ represents the parameters of a pretrained feature extractor. L_{syn} and L_{ref} are determined by T_{syn} and T_{ref}, respectively.

Using $\tilde{S}$ and $\tilde{R}$, we compute similarities between each pair of embeddings and represent them in matrix form. The similarity matrix $M \in \mathbb{R}^{L_{\mathrm{syn}} \times L_{\mathrm{ref}}}$ is defined by cosine similarity for each element (i, j) as:

$$M_{ij} = \mathrm{sim}(\tilde{\boldsymbol{s}}_i, \tilde{\boldsymbol{r}}_j) = \frac{\tilde{\boldsymbol{s}}_i \cdot \tilde{\boldsymbol{r}}_j}{\|\tilde{\boldsymbol{s}}_i\| \cdot \|\tilde{\boldsymbol{r}}_j\|} \tag{1}$$

3.2. Score calculation

Scores are calculated based on the similarity matrix. We first apply a method based on max-norm, as used in BERTScore and SpeechBERTScore. Then, considering non-locality of environmental sounds, we propose a method based on the p-norm. Figure 2 shows the computation.

3.2.1. Computation based on maximum scores

We compute precision, recall, and F1 score from the similarity matrix. Precision is the average of the maximum similarity for each frame in the synthesized embeddings, representing how well the synthesized audio covers the reference. Recall is the average of the maximum similarity for each frame in the reference embeddings, indicating how well the reference is covered by the synthesized one. The harmonic mean of these scores gives the F1 score. These are calculated as

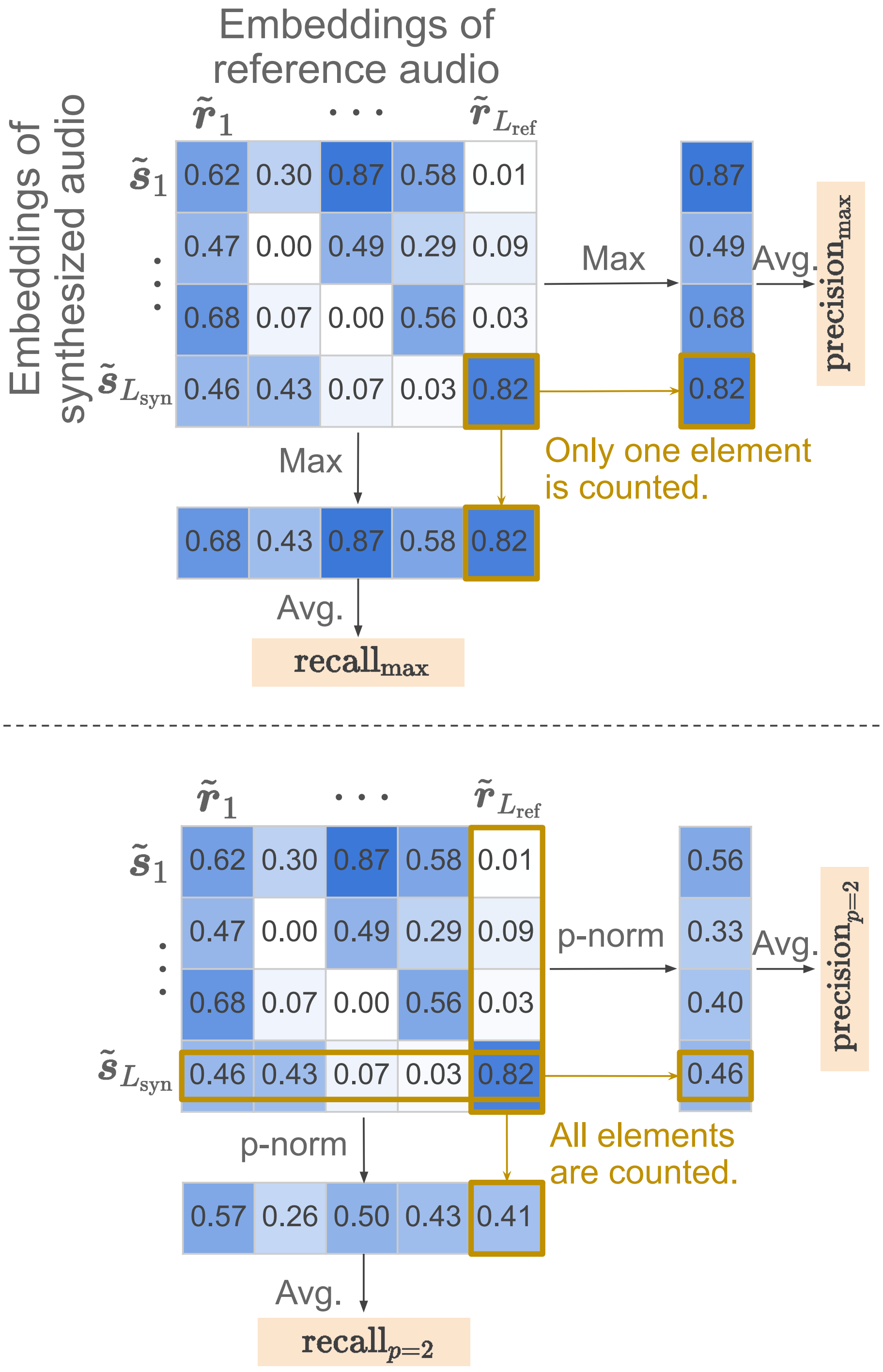
$$\mathrm{precision}_{\mathrm{max}} = \frac{1}{L_{\mathrm{syn}}} \sum_{i=1}^{L_{\mathrm{syn}}} \max_{j=1,\ldots,L_{\mathrm{ref}}} M_{ij} \tag{2}$$

$$\mathrm{recall}_{\mathrm{max}} = \frac{1}{L_{\mathrm{ref}}} \sum_{j=1}^{L_{\mathrm{ref}}} \max_{i=1,\ldots,L_{\mathrm{syn}}} M_{ij} \tag{3}$$

$$\mathrm{F1}_{\mathrm{max}} = 2 \times \frac{\mathrm{precision}_{\mathrm{max}} \times \mathrm{recall}_{\mathrm{max}}}{\mathrm{precision}_{\mathrm{max}} + \mathrm{recall}_{\mathrm{max}}} \tag{4}$$

These scoring methods use the ∞-norm (max-norm), which assumes high-similarity embeddings are temporally localized. This assumption generally holds for natural language and speech, where phrases and segmental features are temporally bounded.

However, this assumption may not hold for environmental sounds. For example, a gunshot sound is a localized, instantaneous sound where embeddings cluster temporally, making locality assumptions valid. In contrast, unstructured, continuous sounds like a babbling brook may have embeddings spread across time, violating this assumption. Therefore, a scoring method that can capture such non-local characteristics is needed.


Embeddings of reference audio
Embeddings of synthesized audio
$\tilde{\boldsymbol{r}}_1$
$\cdots$
$\tilde{\boldsymbol{r}}_{L_{\mathrm{ref}}}$
$\tilde{\boldsymbol{s}}_1$
$\vdots$
$\tilde{\boldsymbol{s}}_{L_{\mathrm{syn}}}$
Max
Avg.
$\mathbf{precision}_{\mathrm{max}}$
Only one element is counted.
Max
Avg.
$\mathbf{recall}_{\mathrm{max}}$
$\tilde{\boldsymbol{r}}_1$
$\cdots$
$\tilde{\boldsymbol{r}}_{L_{\mathrm{ref}}}$
$\tilde{\boldsymbol{s}}_1$
$\vdots$
$\tilde{\boldsymbol{s}}_{L_{\mathrm{syn}}}$
p-norm
Avg.
$\mathbf{precision}_{p=2}$
All elements are counted.
p-norm
Avg.
$\mathbf{recall}_{p=2}$

$\tilde{\boldsymbol{s}}_1$	0.62	0.30	0.87	0.58	0.01	0.87
	0.47	0.00	0.49	0.29	0.09	0.49
	0.68	0.07	0.00	0.56	0.03	0.68
$\tilde{\boldsymbol{s}}_{L_{\mathrm{syn}}}$	0.46	0.43	0.07	0.03	0.82	0.82
	0.68	0.43	0.87	0.58	0.82	

$\tilde{\boldsymbol{s}}_1$	0.62	0.30	0.87	0.58	0.01	0.56
	0.47	0.00	0.49	0.29	0.09	0.33
	0.68	0.07	0.00	0.56	0.03	0.40
$\tilde{\boldsymbol{s}}_{L_{\mathrm{syn}}}$	0.46	0.43	0.07	0.03	0.82	0.46
	0.57	0.26	0.50	0.43	0.41	

Figure 2: Similarity computation in our method. Max-norm (top) assumes locality, and p-norm (bottom) reflects the non-locality.

3.2.2. Computation based on p-norm

Replacing the max-norm with the p-norm, we define the following scores:

$$\text{precision}_p = \frac{1}{L_{\text{syn}}} \sum_{i=1}^{L_{\text{syn}}} \left(\frac{1}{L_{\text{ref}}} \sum_{j=1}^{L_{\text{ref}}} M_{ij}^p \right)^{1/p} \tag{5}$$

$$\text{recall}_p = \frac{1}{L_{\text{ref}}} \sum_{j=1}^{L_{\text{ref}}} \left(\frac{1}{L_{\text{syn}}} \sum_{i=1}^{L_{\text{syn}}} M_{ij}^p \right)^{1/p} \tag{6}$$

As shown in the bottom of Figure 2, these are computed using the p-norm. When $p = 1$, the metrics are simple averages and measure overall (non-local) similarity across time. As p increases, more weight is placed on local peaks, capturing locality. When $p \to \infty$, the score is equivalent to the max-norm.

To balance local and non-local similarity, we introduce the following interpolation between max-based and p-norm-based scores:

$$\text{precision}_{\lambda,p} = \lambda \cdot \text{precision}_{\max} + (1 - \lambda) \cdot \text{precision}_p \tag{7}$$

$$\text{recall}_{\lambda,p} = \lambda \cdot \text{recall}_{\max} + (1 - \lambda) \cdot \text{recall}_p \tag{8}$$

$$\text{F1}_{\lambda,p} = 2 \times \frac{\text{precision}_{\lambda,p} \times \text{recall}_{\lambda,p}}{\text{precision}_{\lambda,p} + \text{recall}_{\lambda,p}} \tag{9}$$

$\lambda \in [0, 1]$ is a hyperparameter for interpolation. When $\lambda = 1$ or $p \to \infty$, the score is equivalent to the max-norm.

4. Experimental evaluation

To evaluate our method, we computed the correlation between the objective scores calculated by our method and the subjective scores.

4.1. Experimental conditions

Dataset. We used two datasets: the PAM test set Deshmukh et al. (2024) and a test set newly developed using the Clotho dataset Drossos et al. (2020), named *Clotho OVL-REL test set*. The former set is used to optimize p, λ, and feature extractors. The latter is used to evaluate the performance of the proposed metric with the optimized configurations. The latter set has no data leakage; there is no overlap with the PAM test set and also no overlap with any of training data for feature extractors. The latter dataset will be released on the project page.

- **PAM test set**: This dataset Deshmukh et al. (2024) consists of synthesized and reference audio, English text, and subjective scores. This set includes 100 pairs of natural audio (reference) randomly extracted from AudioCaps Kim et al. (2019) and their captions, along with 400 synthesized audio by MelDiffusion, AudioLDM 2[3] Liu

[3] https://github.com/haoheliu/AudioLDM2

et al. (2024), AudioLDM-Large[4] Liu et al. (2023), and AudioGen-base[5] Kreuk et al. (2023). Each sample of both reference and synthesized audio is annotated with 5-point MOS scores for OVL and REL. Each MOS score is the average of scores given by 10 different raters. We excluded 17 reference audio samples with REL subjective scores less than 3.5 and the corresponding 17×4 synthesized audio samples. This exclusion was made to ensure the proposed method accurately estimates scores correlated to the REL subjective scores, which assume a strong relation between the reference audio and text. The duration of each sample is 5 seconds for synthesized audio and 10 seconds for reference audio. All audio samples were downsampled at 16 kHz.

- **Clotho OVL-REL test set**: The original Clotho dataset Drossos et al. (2020) consists of natural environmental sounds, each paired with five human-written captions. We used the natural sounds as the reference audio and adopted the first caption (`Caption1`) as the corresponding text. From the original Clotho test set, we selected 100 samples. To maintain diversity among the selected samples in terms of both text content and acoustic features, we used the diversity-based core-set selection algorithm Seki et al. (2024). In this algorithm, samples are selected to diversify the total distance of text (BERT) and audio (PANNs Kong et al. (2020)) features among samples.

  For each caption, we used four different synthesis systems to generate a total of 4×100 audio samples. The TTA systems were almost the same as the PAM test set except for MelDiffusion. Since MelDiffusion is not open-sourced, we used TangoFlux[6] Hung et al. (2024) instead.

  In addition, we used a crowdsourcing service[7] to collect subjective evaluation scores. MOS scores were collected following the same instructions as the PAM test set. To remove low-quality data, we excluded 19 reference audio samples with REL scores less than 3.5 and the corresponding 19×4 synthesized audio samples.

  All audio samples were downsampled at 16 kHz. The duration of each synthesized sample was aligned with that of its corresponding reference audio, which ranged from approximately 15 to 30 seconds.

Feature extractors. We used the following three pretrained models as feature extractors for the proposed method.

- **BYOL-A Niizumi et al. (2023)**[8]: A model based on convolutional neural networks (CNN) LeCun et al. (1989). We used the latest v2. Frame-level embeddings ("local") its channel-flattened feature ("global"), and their concatenation along the feature dimension ("local+global") were used.

[4] https://github.com/haoheliu/AudioLDM
[5] https://github.com/facebookresearch/audiocraft
[6] https://huggingface.co/spaces/declare-lab/tangoFlux
[7] https://www.prolific.com
[8] https://github.com/nttcslab/byol-a/blob/master/v2/AudioNTT2022-BYOLA-64x96d2048.pth

- **ATST-Frame Li et al. (2024)**[9]: A 13-layer Transformer Vaswani et al. (2017)-based model. Features from each of the 1st–13th layers were used. The `ATST-Frame-base` model was used.
- **AST Gong et al. (2021)**[10]: A 13-layer Transformer Vaswani et al. (2017)-based model finetuned on environmental sound classification. Features from each of the 1st through 13th layers were used. We used the fine-tuned model with Full AudioSet, 10 t-stride, 10 f-stride, and weight averaging (0.459 mAP).

Comparison methods. As comparison methods under the same conditions (Table 1), we used MCD Kubichek (1993) and WARP-Q Jassim et al. (2021), which do not require training and use reference audio. Although FAD Kilgour et al. (2019) also falls under the same condition, it is excluded from comparison since it is calculated over multiple samples, unlike the proposed method, MCD, and WARP-Q which are computed per sample. Additionally, as methods under different conditions, we also included CLAPScore Xiao et al. (2024) and PAM Deshmukh et al. (2024).

Evaluation method. For each of the OVL and REL subjective scores, we computed the linear correlation coefficient (LCC) and Spearman's rank correlation coefficient (SRCC) with the objective evaluation metrics. These coefficients were computed across all samples in the test set.

4.2. Result 1: Feature extractors with precision, recall, and F1

Comparison of feature extractors.

	OVL		REL	
	LCC	SRCC	LCC	SRCC
AST, 13th layer, $\mathrm{F1}_{\mathrm{max}}$	**0.426**	**0.395**	**0.339**	**0.317**
ATST-Frame, 10th layer, $\mathrm{recall}_{\mathrm{max}}$	0.366	0.362	0.256	0.250
BYOL-A, global feature, $\mathrm{recall}_{\mathrm{max}}$	0.055	0.045	0.116	0.121

Table 2: Comparison for each feature extractor. The score calculation based on the maximum value and the features extracted use the best settings. (the Clotho OVL-REL test set)

We compared the correlation between scores calculated from each layer of the feature extractors and the subjective evaluation scores. The scores were computed using the max-based method (Equations (2)–(4)). Results are shown in Figure 3. AST and ATST-Frame showed higher correlation in later layers, indicating that later layers capture contextual information relevant for environmental sounds, which benefits AudioBERTScore. For BYOL-A, the OVL score is higher for the local embedding, while the REL score is higher for global or local+global, likely because local captures acoustic features near the input layer, while global retains contextual information.

[9] https://github.com/Audio-WestlakeU/audiossl/blob/main/audiossl/methods/ATST-Frame/README.md

[10] https://github.com/YuanGongND/ast/blob/master/pretrained_models/README.md

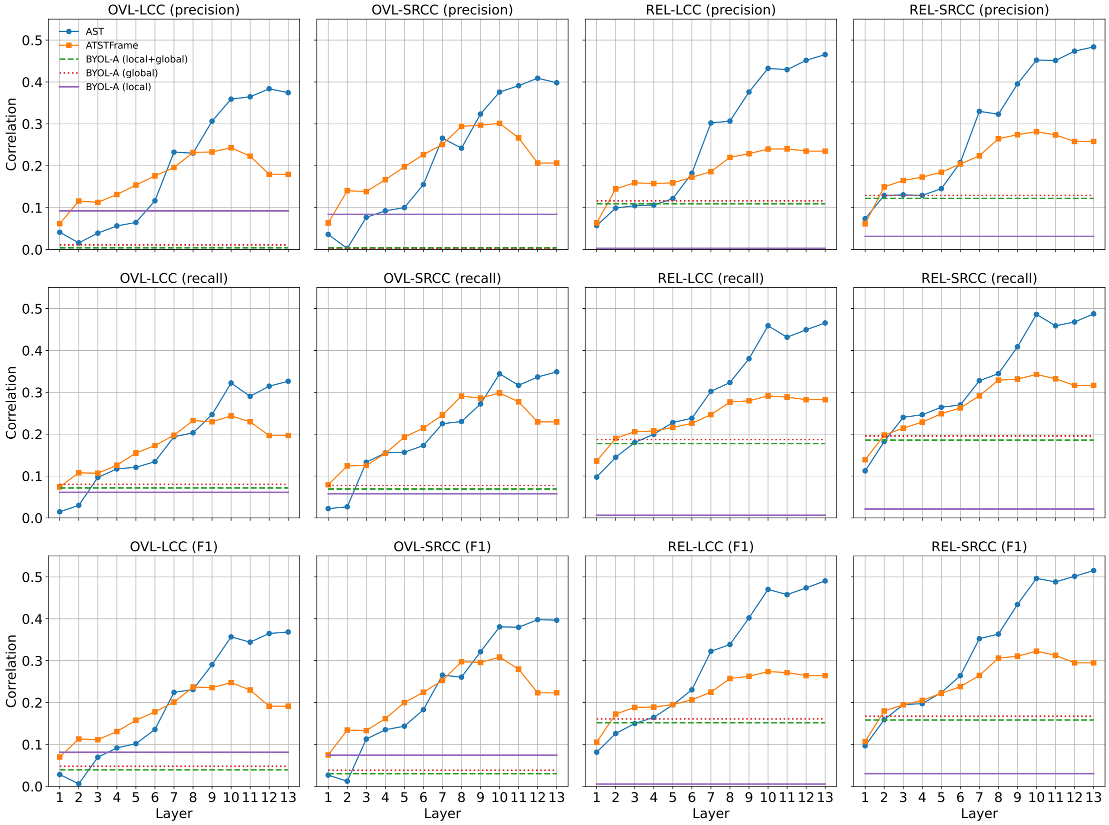


Figure 3: Correlation under several settings of feature extractors and similarity computation. (The PAM test set)

Comparison of precision, recall, and F1. Figure 3 indicates that both AST and ATST-Frame show a tendency for higher recall than precision. In AST, precision rapidly increases around layers 6–7, and recall increases sharply around layers 8–10. Further investigation is needed to explain these patterns.

Comparison of best configurations. Table 2 shows the results for the best configurations for each feature extractor, chosen based on average high correlation scores in Figure 3. AST uses $\mathrm{F1}_{\mathrm{max}}$ from the 13th layer, ATST-Frame uses $\mathrm{recall}_{\mathrm{max}}$ from the 10th layer, and BYOL-A uses $\mathrm{recall}_{\mathrm{max}}$ from the global layer.

The Transformer-based models, AST and ATST-Frame, significantly outperform the CNN-based BYOL-A, indicating the effectiveness of contextual information extraction with Transformers Peng et al. (2023). Furthermore, AST's superior performance suggests that fine-tuning for environmental sound classification contributed to the improvement.

4.3. Result 2: Max-norm vs. p-norm

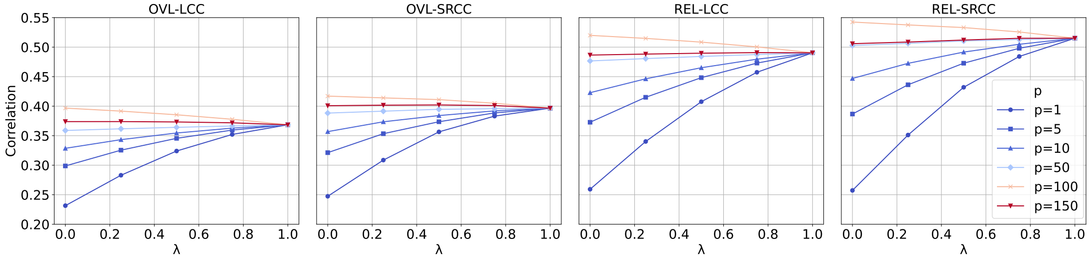


Figure 4: Correlation using p-norm-based calculation with various p and λ. (The PAM test set)

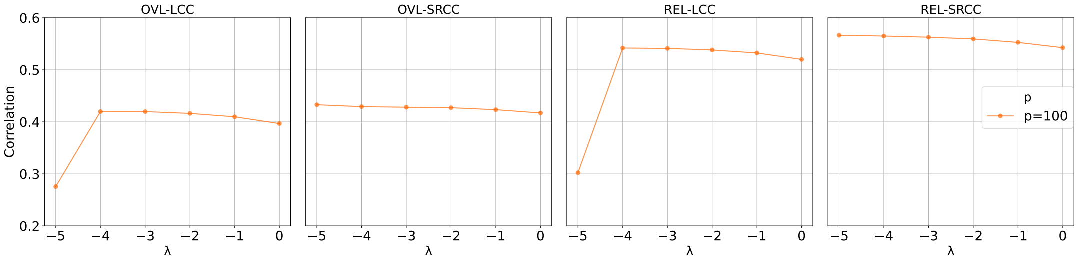


Figure 5: Correlation using $p = 100$ and negative values of λ. (The PAM test set)

Effect of p and λ. We investigated the performance of the p-norm based score (Equations (7)–(9)) with various values of p and λ. Results using the F1 score with the 13th layer of AST (the best performer in Table 2) are shown in Figure 4. Performance peaked at $p = 100$, especially at $\lambda = 0$. Interestingly, $p = 106$ slightly outperformed $p = 100$.

Investigation of negative λ. From the trend in Figure 4, correlation increases as λ decreases, even continuing beyond $\lambda = 0$. Thus, we explored negative λ values to assess potential further improvements. Figure 5 shows correlation improves until $\lambda = -4$, then drops at $\lambda = -5$. We hypothesize that both localized and non-localized similarity contribute positively when the p-norm weight increases and maximum-based score grows in magnitude.

4.4. Result 3: Comparison with other objective evaluation metrics

We compared the performance of the proposed method with existing objective evaluation metrics. Results are shown in Table 3, using the best-performing configurations with $p = 106$ and $\lambda = -3.5$ as derived from Figure 5. While not direct competitors in terms of the evaluation methods used, PAM and CLAPScore are included for reference.

The proposed method significantly outperformed MCD and WARP-Q in both REL and OVL correlations, showing its usefulness as a training-free, reference-based evaluation metric. Comparing variants in the proposed metric, introducing p-norm or negative λ decreases OVL-LCC (SRCC) but increases REL-LCC (SRCC). This suggests that the use of non-localized information inhibit to capture sound quality, but enhance to capture contextual information.

Lastly, we compare the proposed method with PAM and CLAPScore. The proposed method achieved the highest correlations in REL and OVL-LCC, supporting its contribution to objective audio evaluation, whereas PAM showed a slightly higher correlation in OVL-SRCC.

	OVL		REL	
	LCC	SRCC	LCC	SRCC
Compared metrics				
MCD	0.158	0.112	0.059	0.031
WARP-Q	0.027	0.052	0.045	0.013
Proposed AudioBERTScore (AST, 13th layer)				
$\mathrm{F1}_{\max}$	**0.426**	**0.395**	0.339	0.317
$\mathrm{F1}_{\lambda=0,p=106}$	0.420	0.346	**0.392**	0.335
$\mathrm{F1}_{\lambda=-3.5,p=106}$	0.374	0.332	0.375	**0.354**
Other metrics				
PAM	0.403	0.417	0.147	0.143
CLAPScore w/ LAION-CLAP	0.218	0.221	0.361	0.339
CLAPScore w/ MS-CLAP	0.199	0.213	0.130	0.128

Table 3: Evaluation results for each objective evaluation metric. Higher values indicate a stronger correlation with the subjective evaluation scores. (The Clotho OVL-REL test set)

4.5. Result 4: Event-wise performance analysis

To investigate the proposed metric, we decompose scores in Table 3 by each audio category. We assigned one of the top-level categories in the AudioSet Gemmeke et al. (2017) ontology to each sample in the Clotho OVL-REL test set. For the purpose, we mapped descriptive keywords in the original Clotho dataset to the ontology.

			OVL		**REL**	
Category	**n samples**	Metric	**LCC**	**SRCC**	**LCC**	**SRCC**
Animal	64	F1$_{\max}$	0.361	0.277	0.336	0.270
		F1$_{\lambda=-3.5,p=106}$	**0.402**	**0.323**	**0.337**	**0.291**
Channel, environment and background	36	F1$_{\max}$	0.314	0.255	0.138	0.049
		F1$_{\lambda=-3.5,p=106}$	**0.317**	**0.316**	**0.143**	**0.167**
Natural sounds	80	F1$_{\max}$	0.387	0.426	0.431	0.424
		F1$_{\lambda=-3.5,p=106}$	**0.406**	**0.462**	**0.468**	**0.468**
Sounds of things	88	F1$_{\max}$	**0.264**	**0.259**	**0.122**	**0.159**
		F1$_{\lambda=-3.5,p=106}$	0.150	0.135	0.103	0.143
Human sounds	52	F1$_{\max}$	0.603	0.611	0.497	0.516
		F1$_{\lambda=-3.5,p=106}$	**0.637**	**0.640**	**0.511**	**0.531**
Music	4	F1$_{\max}$	0.577	0.632	0.809	0.800
		F1$_{\lambda=-3.5,p=106}$	**0.597**	**0.632**	**0.911**	**0.800**
Overall	324	F1$_{\max}$	**0.426**	**0.395**	0.339	0.317
		F1$_{\lambda=-3.5,p=106}$	0.374	0.332	**0.375**	**0.354**

Table 4: Correlation results per category under two settings (F1$_{\max}$ and F1$_{\lambda=-3.5,p=106}$). Higher values indicate better correlation. (The Clotho OVL-REL test set)

Table 4 shows the results. The tendencies are completely different among "Sounds of things" and other categories. The use of p-norm and negative λ improves scores in all categories other than "Sounds of things." Since those categories include sounds with spectro-temporal patterns, e.g., human speech, the configurations contribute to capture such kinds of sounds. On the other hand, the "Sounds of things" category includes non-stationary and impulsive sounds, such as doors closing or objects colliding, the use p-norm inhibit to capture.

5. Conclusion

In this paper, we proposed an objective evaluation metric for TTA based on the similarity between sequences of synthesized and reference audio embeddings. Evaluation results demonstrated that the proposed method achieved the best performance among unsupervised, audio-reference metrics. Furthermore, it outperformed other metrics, e.g., PAM and CLAPScore. As future work, we plan to explore improved similarity and score computation methods.

Acknowledgments

This research was supported by JSPS KAKENHI 23K18474 and 25K21221, JST FOREST JP23KJ0828, JPNP25006, a project outsourced by the New Energy and Industrial Technology Development Organization (NEDO) ,and Moonshot R&D Grant Number JPMJPS2011.

References

Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 359–370, 1994.

Keunwoo Choi, Jaekwon Im, Laurie M. Heller, Brian McFee, Keisuke Imoto, Yuki Okamoto, Mathieu Lagrange, and Shinnosuke Takamichi. Foley sound synthesis at the DCASE 2023 challenge. In *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, pages 16–20, Tampere, Finland, September 2023.

Soham Deshmukh, Dareen Alharthi, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, and Huaming Wang. PAM: Prompting Audio-Language Models for Audio Quality Assessment. In *Proceedings of Interspeech*, pages 3320–3324, 2024. doi: 10.21437/Interspeech.2024-325.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, Minnesota, 2019. doi: 10.18653/v1/N19-1423.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740, 2020. doi: 10.1109/ICASSP40776.2020.9052990.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Maurice Fréchet. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo*, 22(1):1–72, 1906. doi: 10.1007/BF03018603.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017.

Yuan Gong, Yu-An Chung, and James Glass. AST: Audio spectrogram transformer. In *Proceedings of Interspeech*, pages 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Shengchuan Han and Lisheng Zhang. Dynamic time warping under subsequence. In *Proceedings of the 4th International Conference on Information Science, Electrical, and Automation Engineering*, volume 12257, page 122571X, 2022. doi: 10.1117/12.2640305.

John H. L. Hansen and Bryan L. Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP)*, 1998.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. tangoflux: Super fast and faithful text-to-audio generation with flow matching and CLAP-ranked preference optimization, 2024.

Wissam A. Jassim, Jan Skoglund, Michael Chinen, and Andrew Hines. WARP-Q: Quality prediction for generative neural speech codecs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 401–405, 2021. doi: 10.1109/ICASSP39728.2021.9414901.

Yusuke Kanamori, Yuki Okamoto, Taisei Takano, Shinnosuke Takamichi, Yuki Saito, and Hiroshi Saruwatari. Relate: Subjective evaluation dataset for automatic evaluation of relevance between text and audio. In *Proceedings of Interspeech*, August 2025.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech*, pages 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee 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 (NAACL-HLT)*, 2019.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. doi: 10.1109/TASLP.2020.3030497.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually guided audio generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

R. Kubichek. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, volume 1, pages 125–128, 1993. doi: 10.1109/PACRIM.1993.407206.

Yann LeCun, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541.

Xian Li, Nian Shao, and Xiaofei Li. Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1336–1351, 2024. doi: 10.1109/TASLP.2024.3352248.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 21450–21474, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. doi: 10.1109/TASLP.2024.3399607.

Thomas Marrinan, Pakeeza Akram, Oli Gurmessa, and Anthony Shishkin. Leveraging AI to generate audio for user-generated content in video games, 2024.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. BYOL for Audio: Exploring pre-trained general-purpose audio representations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:137–151, 2023. doi: 10.1109/TASLP.2022.3221007.

Yuki Okamoto, Keisuke Imoto, Shinnosuke Takamichi, Takahiro Fukumori, and Yoichi Yamashita. How should we evaluate synthesized environmental sounds. In *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 307–312, 2022. doi: 10.23919/APSIPAASC55919.2022.9980187.

Zhiliang Peng, Zonghao Guo, Wei Huang, Yaowei Wang, Lingxi Xie, Jianbin Jiao, Qi Tian, and Qixiang Ye. Conformer: Local features coupling global representations for recognition and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9454–9468, 2023. doi: 10.1109/TPAMI.2023.3243048.

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. SpeechBERTScore: Reference-aware automatic evaluation of speech generation leveraging nlp evaluation metrics. In *Proceedings of Interspeech*, pages 4943–4947, 2024.

Kentaro Seki, Shinnosuke Takamichi, Takaaki Saeki, and Hiroshi Saruwatari. Diversity-based core-set selection for text-to-speech with linguistic and acoustic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12351–12355, 2024. doi: 10.1109/ICASSP48485.2024.10448068.

Taisei Takano, Yuki Okamoto, Yusuke Kanamori, Yuki Saito, Ryotaro Nagase, and Hiroshi Saruwatari. Human-CLAP: Human-perception-based contrastive language-audio pretraining. In *Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 131–136, October 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Feiyang Xiao, Jian Guan, Qiaoxi Zhu, Xubo Liu, Wenbo Wang, Shuhan Qi, Kejia Zhang, Jianyuan Sun, and Wenwu Wang. A reference-free metric for language-queried audio source separation using contrastive language-audio pretraining. In *Proceedings of the DCASE Workshop*, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.