RELATE: Subjective evaluation dataset for automatic evaluation of relevance between text and audio

Yusuke Kanamori[1], *Yuki Okamoto*[1], *Taisei Takano*[1], *Shinnosuke Takamichi*[2,1], *Yuki Saito*[1], *Hiroshi Saruwatari*[1]

[1]The University of Tokyo, Japan
[2]Keio University, Japan

kanamori-yusuke796@g.ecc.u-tokyo.ac.jp, y-okamoto@ieee.org

Abstract

In text-to-audio (TTA) research, the relevance between input text and output audio is an important evaluation aspect. Traditionally, it has been evaluated from both subjective and objective perspectives. However, subjective evaluation is costly in terms of money and time, and objective evaluation is unclear regarding the correlation to subjective evaluation scores. In this study, we construct *RELATE,* an open-sourced dataset that subjectively evaluates the relevance. Also, we benchmark a model for automatically predicting the subjective evaluation score from synthesized audio. Our model outperforms a conventional CLAPScore model, and that trend extends to many sound categories.

Index Terms: text-to-audio, human evaluation, CLAPScore, environmental sound synthesis

1. Introduction

Research on text-to-audio (TTA), which is a technology to automatically synthesize an audio sample from text, such as "a dog barking behind a human speech," is attracting attention [1]. TTA has much potential, such as generating background sounds and sound effects for media contents [2] and creating audio environments in virtual reality.

TTA is evaluated from both subjective and objective perspectives. Subjective evaluation of TTA can be broadly divided into audio quality and relevance. The former is an evaluation of whether the synthesized audio samples are of high quality, while the latter is of assessment of the extent to which the synthesized audio samples reflect the content of the input text. Recent TTA models synthesize high-quality audio samples but often omit content from the input text [3]. Therefore, we focus on the subjective evaluation of relevance. On the other hand, this requires time and money, and it is impossible to compare scores in different listening tests. Therefore, realizing an objective evaluation metric that is highly correlated to human subjectivity is an important research topic.

This problem is not limited to TTA but is prevalent in various generative tasks. To address this issue, supervised machine learning methods have been proposed in speech synthesis and image generation to automatically predict subjective evaluation scores from synthesized outputs [4, 5]. These methods train a machine learning model from paired data of synthesized outputs and subjective evaluation scores, enabling the prediction of subjective evaluation scores for unseen synthesized outputs. This approach holds promise for application in TTA to simplify evaluation. Furthermore, the subjective evaluation scores of synthesized audio tend to exhibit significant variance among evaluators and synthesized outputs [6, 3]. When predicting subjective evaluation scores for TTA, it is necessary to investigate and consider the influence of listener, audio, and text attributes.

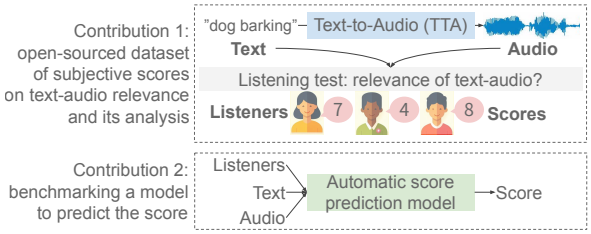

Figure 1: *Overview of this study.*

The contributions of our paper are illustrated in Figure 1. We construct an open-source dataset, which is called *RELATE* (**REL**evance score on **A**udio and **TE**xt), consisting of synthesized audio samples and relevance scores[1]. The collection of scores for synthesized and original audio samples can be expected to be used as a screening method when large amounts of data for TTA are obtained from the internet. The dataset covers three attributes regarding 1) listener, 2) synthesized audio, and 3) text, and in this study, we investigate the influence of these attributes on subjective evaluation scores. Furthermore, we train a model using the constructed dataset to predict the relevance between text and audio samples and conduct benchmark analysis. The results show that our model outperforms CLAPScore [7], and that trend extends to many sound categories.

2. Related work

In the field of speech synthesis, a method for predicting subjective scores on the naturalness of synthesized speech has been proposed. The VoiceMOS Challenge [4], which is an international competition to assess the performance of automatic subjective score prediction, has also been held with a shared dataset of synthesized speech and subjective scores. Some prediction models have been proposed [8], and the self-supervised learning (SSL) models (e.g., wav2vec 2.0 [9], WavLM [10]) are often used as a module to extract speech features for the subjective score prediction. In this study, we make use of these ideas and construct a shared dataset for TTA, as well as build an SSL-based benchmark model. Furthermore, we analyze the factors contributing to the variance in subjective evaluation scores.

In the evaluation of TTA, some objective evaluation metrics have been proposed for both overall audio quality (OVL) and relevance to the text input (REL), and their correlation with subjective evaluation scores has been investigated. Regarding OVL, Deshmukh et al. proposed a prediction method using CLAP [11]. Similarly, Tjandra et al.'s method predicted the val-

[1]https://github.com/sarulab-speech/RELATE

Table 1: *Questionnaire for listener attributes.*

ID	Question	Options
Q01	Age	$\leq$20, 21–30, ..., 61$\leq$
Q02	Gender	M, F, NBi
Q03	How many times have you participated in ratings of audio samples?	0, 1, ..., 5$\leq$
Q04	When did you last participate in other ratings of audio samples?	Never, $\leq$1 month, ...
Q05	On average, how many times have you heard an audio repeatedly?	1, 2, ..., 5$\leq$
Q06	What type of audio device did you use?	Headphone, Earphone, Others
Q07	Was the surrounding environment quiet during the ratings of audio samples?	Quiet, ..., Noisy
Q08	How difficult were the evaluations?	Easy, ..., Difficult
Q09	Do/did you work in the field of speech or audio technology?	Yes, No
Q10	Nationality	EU, NA, ...
Q11	Mother country	EU, NA, ...
Q12	Place of residence	EU, NA, ...

ues of subdivided evaluation axes of OVL [2]. While these studies can predict evaluation scores that correlate to some extent with the subjective evaluation scores of OVL, they cannot be used for the evaluation of REL, which indicates how well the content of the input text is reflected.

Regarding REL, Huang et al. proposed an unsupervised method called CLAPScore [7]. It calculates the cosine similarity between the input text and the synthesized audio using a pretrained CLAP model. It is unclear what the correlation between the CLAPScore and the REL subjective evaluation scores. Although there is a difference between unsupervised and supervised learning, this method should be compared with our constructed benchmark.

3. Creation of dataset

3.1. Overview of dataset

Our dataset consists of the following contents.

- **Text–audio pairs.** Both original and synthetic audio samples are included.
- **Subjective evaluation scores.** Three metrics of 11-point scores for each text–audio referring to the DCASE 2024 Challenge Task 7[3].
  - **REL score.** The overall relevance of the text and audio.
  - **Inclusion of sound event (IS) score.** The extent to which the sound events described in the text are included in the audio.
  - **Order of sound event (OS) score.** The degree of matching between the time series of sound events described in the text and the audio.
- **Listener attributes.** Listener ID, age, gender, nationality, birthplace, residence, and experience of audio evaluation. Table 1 shows the questions and options of listener attributes.

3.2. Collecting subjective evaluation scores

We collected subjective scores for each text–audio sample. For original audio samples, 1,000 pairs, including 500 pairs with words indicating the order of sound occurrence, i.e., "before," "after," 'then," or "followed by," selected on the basis of the conventional study [12] and 500 pairs without such words, were randomly selected from each of the AudioCaps training and test datasets [13]. Synthesized audio samples were obtained using the open-sourced pretrained TTA models: AudioLDM [14], AudioLDM2 [15], Tango [16], and Tango2 [17]. The texts of original audio samples are used as input to these TTA models. For each text, two synthesis models are selected and synthesized. Furthermore, referring to [18], we presented explanations for each score shown in Table 2 to minimize score variations among listeners. Listeners were presented with the audio and text, and they answered each metric on an 11-point scale.

Table 2: *Explanations of each score in subjective evaluations.*

Metric	Score	Instruction
REL	0	Does not match at all.
	2	Has significant discrepancies.
	5	Has several minor discrepancies.
	8	Has a few minor discrepancies.
	10	Matched exactly.
IS	0	All sound events are clearly missing.
	2	Most of the sound events seem to be missing.
	5	About half of the sound events seem to be missing.
	8	Most of the sound events seem to be included.
	10	All sound events are clearly included.
OS	0	All sound events in the audio clearly occurred in the wrong order.
	2	Most sound events in the audio occurred in the wrong order.
	5	About half of the sound events in the audio occurred in the correct order.
	8	Most sound events in the audio occurred in the correct order.
	10	All sound events in the audio clearly occurred in the correct order.

Table 3: *Statistics of RELATE dataset.*

	REL		IS		OS	
	Train	Test	Train	Test	Train	Test
Evaluations	9,963	7,797	7,641	5,865	4,017	2,943
Audio–text pairs	2,862	2,598	2,649	2,334	1,281	1,185
Audio duration [s]	28,806	26,129	26,654	23,476	12,880	11901
Listeners	1,085	873	864	635	714	525

Each listener answered each metric after being presented with a text–audio pair. Note that the REL score was obtained through an independent experiment, separate from the IS and OS score collection described later. In collecting IS and OS scores, the same listeners conducted both evaluations.

In annotating IS scores, some text samples included words indicating the order of sound events, such as "before" and "after." Listeners were instructed to disregard the order of occurrence as long as the sound events described in the text were present in the audio. In annotating the OS score, listeners were instructed to score a "0" if the text did not represent the order of sound occurrence and these scores were subsequently omitted from our OS dataset.

3.3. Screening

To conduct screening to ensure data quality and the result of evaluation, each evaluation set was intentionally designed to include samples with low relevance between text and audio. Original audio in AudioCaps is annotated with multiple sound event labels in addition to the text describing the audio. Audio samples with mismatched sound event labels were randomly selected from the dataset, and these audio samples and texts were paired and set as anchors.

Based on the answers for the anchors in the evaluation, we screened listeners for training and test sets. For the training set, we excluded listeners whose average anchor rating was "2" or higher. For the test set, we excluded listeners whose average anchor rating was "1" or higher to ensure high quality of collected scores. Table 3 shows the statistics of our dataset after the screening.

[2] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov, and others, "Meta Audiobox Aesthetics: Unified Automatic Quality Assessment for Speech, Music, and Sound," arXiv preprint arXiv:2502.05139, 2025.

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

Table 4: *Statistical significances ($p < 0.05$) among items and of interaction between items of natural/synthetic audio samples for REL scores. Check mark shows significant differences and interaction.*

Factor	Among items	Interaction
Sound event labels in text		
Number of event labels		
Number of top-level sound categories	✓	✓
Sounds belonging to a category vs. sounds in other categories		
Human sounds		✓
Animal	✓	✓
Natural sounds		✓
Music	✓	
Sounds of things		
Source-ambiguous sounds		
Channel, environment and background		
Speech		✓
Text complexity		
Number of words	✓	✓
w/ temporal preposition vs. w/o	✓	✓
Flesch Reading Ease	✓	

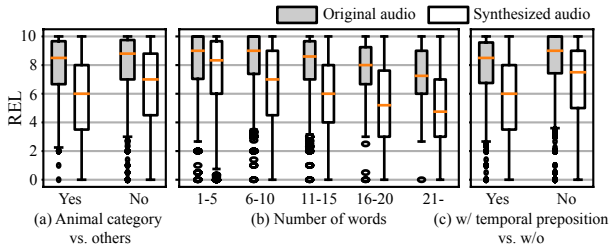

Figure 2: *Boxplots of each aspect*

Table 5: *Distribution of training, validation, and test data.*

AudioCaps	Train	Test	
RELATE dataset	Train	Validation	Test
Evaluations	9,963	3,897	3,900
Audio–text pairs	2,862	1,287	1,311
Audio duration [s]	28,806	12,960	13,169
Listeners	1,085	712	726

4. Analysis of dataset

REL is the most commonly used metric in three metrics, and we have conducted an analysis of REL. We focused on the REL and analyzed the subjective evaluation dataset in the following aspects: audio attribute and text attribute. Furthermore, we investigated whether these trends differ between original and synthesized audio samples. A stricter screening than those in the dataset creation process was performed in this analysis to reduce noise in the evaluation and to acquire meaningful tendency. In addition to excluding listeners with an average anchor rating of "2" or higher, we added the exclusion of listeners with an average original audio rating of "6" or lower to avoid listeners who rated everything low. Then, we removed listeners whose ratings had the lowest 5% entropy of all listeners.

We conducted nonparametric tests: Mann–Whitney U test [19] for two-group means, Kruskal–Wallis test [20] for 3+-group means, and Steel–Dwass test [21] for multiple comparison. In addition, aligned rank transform (ART) analysis of variance (ANOVA) [22], a method for nonparametric data, was used to examine the interactions of two factors.

In TTA, audio attributes and text attributes are important. Original audio samples of AudioCaps have labels indicating the type of sound. This label is reflected in the text, which in turn is reflected in the synthesized audio. Then, those labels can be applied to synthesized audio samples as audio attributes. Text attributes are characterized by their complexity. We examined our dataset using audio attributes and text attributes in the following three aspects.

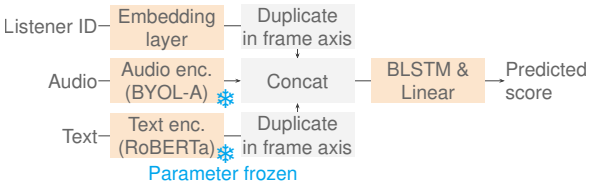

Figure 3: *Overview of proposed REL prediction model*

Table 6: *Results of subjective evaluation prediction.*

Method	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑
CLAPScore w/ MS-CLAP	0.159	0.208	0.181	0.124
CLAPScore w/ LAION-CLAP	0.082	0.375	0.351	0.242
Ours	0.073	**0.385**	**0.383**	**0.265**
Ours w/o CBL	**0.069**	0.377	0.374	0.259

- **Sound event labels.** The number of sound event labels and the number of top-level categories.
- **Sounds belonging to a category vs. sounds in other categories.** Sounds belong to a top-level category and others, e.g., human sounds vs. others.
- **Text complexity.** Number of words in the text, with or without temporal prepositions, and the Flesch Reading Ease score [23], which indicates the readability of English text.

4.1. Analysis from sound event label

Upper part of Table 4 shows which audio factors had statistically significant differences and which audio factors interacted with original/synthetic audio samples. It shows that top-level categories affect the evaluation score. It is assumed that the kind of top-level categories affects both the difficulty of human evaluation and that of synthesis. Specifically, belonging to the "Animal" category shows both of statistically significant differences and interaction. Figure 2(a). shows that synthesized audio in the "Animal" category had the lower score than in others. It is assumed that synthetic models have difficulty in synthesizing animal sounds.

4.2. Analysis from texts

Lower part of table 4 shows which text factors had statistically significant differences and which text factors interacted with original/synthetic audio samples. In particular, number of words and inclusion of temporal preposition show both of statistically significant differences and interaction. Figure 2(b) shows that the score decreases as the number of words increases, and interaction effects indicate that the more words there are, the less successful the synthesis is. This can be thought of as the greater the number of words, the more difficult the subjective evaluation or synthesis becomes. Figure 2(c) shows that the score decreases for texts that include time-series information. The interaction effects indicate that the synthesis is not successful when time-series information is included. It is inferred that the synthetic models are not trained with attention to time-series information.

5. Benchmarking prediction model

5.1. Model architecture

We trained a model to predict the REL score between audio and text. Figure 3 shows the model. The audio $\boldsymbol{x}$ and text w are

Table 7: *Results of subjective evaluation prediction for each top-level category.*

Method	Human sounds	Animal	Natural sounds	Music	Sounds of things	Source-ambiguous sounds	Channel, environment and background	Speech
CLAPScore w/ MS-CLAP	0.181	0.254	0.232	0.310	0.126	0.339	0.170	0.174
CLAPScore w/ LAION-CLAP	0.349	**0.438**	0.311	0.186	0.266	**0.452**	0.338	0.339
Ours	**0.435**	0.411	**0.353**	**0.599**	**0.380**	0.447	**0.462**	**0.445**

input into the pretrained audio and text encoders, respectively: $\boldsymbol{V} = \mathsf{AudioEnc}(\boldsymbol{x}) \in \mathbb{R}^{F \times T}$, $\boldsymbol{o} = \mathsf{TextEnc}(w) \in \mathbb{R}^{D}$. Here, F, T, and D denote the number of dimensions of audio features, time length, and number of dimensions of text features, respectively. For audio and text encoders, we used pre-trained BYOL-A [24] and RoBERTa[4], respectively. Then, $\boldsymbol{o}$ and C-dimension listener-embedding vector $\boldsymbol{l} \in \mathbb{R}^{C}$ are duplicated in the time direction. Listener embeddings enhance prediction accuracy by modeling individual listener preferences [25]. The feature $\boldsymbol{M} \in \mathbb{R}^{(F+C+D) \times T}$ obtained by concatenating the obtained feature sequences $\boldsymbol{V}$ and the temporally duplicated $\boldsymbol{l}, \boldsymbol{o}$ in the dimension direction is input to the bidirectional long short–term memory (BLSTM) [26]. Finally, by passing the output $\boldsymbol{Z}$ from the BLSTM through two linear layers and the activation function ReLU, the REL score is predicted.

5.2. Loss function

The training objective is a weighted sum of two functions, both evaluating the difference between the ground truth and predicted scores: the clipped mean squared error (MSE) loss $\mathcal{L}^{reg}$ [27] and contrastive loss $\mathcal{L}^{con}$ [8]. Class-balanced loss (CBL) [28] was introduced to reduce the influence of data bias. We round up REL scores to the nearest integer. Let $l_y \in [1, 2, ..., 10]$ be the integer-converted version of REL score y, and n_{l_y} be the frequency of class l_y. Then, $E_y = \frac{1-\beta_{cbl}}{1-\beta_{cbl}^{n_{l_y}}}$ is a value that decreases as n_{l_y} increases. Here, $\beta_{cbl} \in [0, 1]$ is a hyperparameter. Then, $\mathcal{CBL}^{reg}$ and $\mathcal{CBL}^{con}$ are obtained by applying E_y to each loss function. The final loss function is obtained by summing the two loss functions with weights β and γ: $\mathcal{L} = \beta\mathcal{CBL}^{reg} + \gamma\mathcal{CBL}^{con}$.

5.3. Experimental setup

Table 5 shows the distribution of training, validation, and test data of REL scores in the RELATE dataset. For model training, we used the training data of REL scores. For validation and evaluation, the test data of REL scores was divided into two subsets so there was no overlap between audio samples and texts. The prediction score was normalized to the range of $[-1, 1]$. In addition to each listener, score prediction was performed for the average listener, whose score is the average of the scores of all listeners [25]. During inference, the model predicted the average listener's score. We empirically chose $\tau = 0.25$ for clipped MSE loss, $\alpha = 0.1$ for contrastive loss, $\beta_{cbl} = 0.99$ for CBL, and $\beta = 1.0$ and $\gamma = 0.5$ for the weights of the two loss functions. The batch size was 12, and gradient accumulation was performed every two steps. Adam [29] ($\beta_1 = 0.9, \beta_2 = 0.999$) was used as the optimizer with an initial learning rate of 2.0×10^{-5}. Also, learning rate scheduling with linear warm-up and linear decay was used. The total number of training steps was 15,000, with up to 4,000 warm-up steps. The optimal model was selected referring to Spearman's rank correlation coefficient (SRCC) calculated from the validation set.

We used CLAPScore for the comparison. As the CLAP model used to calculate CLAPScore, LAION-CLAP [30] and MS-CLAP [31] were used. For the evaluation of the prediction scores of each model, we used MSE, linear correlation coefficient (LCC), SRCC, and Kendall rank correlation coefficient (KTAU) referring to the VoiceMOS Challenge [4].

5.4. Results and discussion

Table 6 shows the results for each evaluation metric. The proposed method outperforms both MS-CLAP and LAION-CLAP in all metrics. It can be seen that the output of the proposed method is closer to the human subjective evaluation value than the similarity score calculated from MS-CLAP and LAION-CLAP. We also found that CBL is effective.

To capture trends in the strengths and weaknesses of subjective evaluation score prediction, SRCCs of prediction and subjective evaluation values were calculated for each top category and compared with the CLAPScore. Table 7 shows the results. Our method outperformed LAION-CLAP and MS-CLAP for all categories except two cases, "Animal" and "Source–ambiguous sounds." In particular, a large difference is observed for "Music," which may be because this method, compared with CLAPScore, reflects the tendency of subjective evaluation to recognize and rate highly music, which has characteristics that are very different from those of other sounds.

Further improvements could be made to the model structure. In the future, experiments using audio encoders other than BYOL-A will be necessary. The quality of the audio encoder is paramount, especially in the field of TTA, which deals with many types of audio. Regarding the text encoder, explicitly encoding the units of sound events, rather than encoding the text as it is, may be helpful for REL prediction. As for the input, we believe that listener attributes help in predicting subjective evaluation scores by providing modeling of the listener.

6. Conclusion

We constructed an open-source dataset consisting of synthesized audio samples and relevance scores. From the analysis result, regarding audio attributes, people have different evaluation tendencies for different types of sound, and there are sounds that the synthesis model does not handle well. For text complexity, we found that the longer the sentence, the lower the evaluation value and the less successful the synthesis. Also, synthesis from texts containing time series was poor. We also built a baseline model for predicting the subjective evaluation score of text–audio relevance in TTA. Our model outperformed the conventional method, CLAPScore, and that trend extended to many sound categories. For future work, we are considering the analysis and prediction of IS and OS scores, the development of a dataset that focuses on audio attributes and text attributes, and the improvement of the prediction model by introducing other encoders or inputs.

[4] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

7. Acknowledgements

The work was supported by JSPS KAKENHI Grant Number 23K24895, 24K23880, 25K21221, JST Moonshot Grant Number JPMJMS2237.

8. References

[1] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 1720–1733, 2023.

[2] D. B. Lloyd, N. Raghuvanshi, and N. K. Govindaraju, "Sound synthesis for impact sounds in video games," in *Symposium on Interactive 3D Graphics and Games*, 2011, pp. 55–62.

[3] J. Lee, M. Tailleur, L. M. Heller, K. Choi, B. M. M. Lagrange, K. Imoto, and Y. Okamoto, "Challenge on sound scene synthesis: Evaluating text-to-audio generation," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[4] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao, "The VoiceMOS Challenge 2024: Beyond speech quality prediction," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, 2024, pp. 803–810.

[5] T. Kou, X. Liu, Z. Zhang, C. Li, H. Wu, X. Min, G. Zhai, and N. Liu, "Subjective-aligned dataset and metric for text-to-video quality assessment," in *Proceedings of the 32nd ACM International Conference on Multimedia*. Association for Computing Machinery, 2024, p. 7793–7802.

[6] Y. Okamoto, K. Imoto, S. Takamichi, R. Nagase, T. Fukumori, and Y. Yamashita, "Environmental sound synthesis from vocal imitations and sound event labels," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 411–415.

[7] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202, 2023, pp. 13 916–13 932.

[8] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022," in *Interspeech 2022*, 2022, pp. 4521–4525.

[9] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[10] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. Wei, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[11] S. Deshmukh, D. Alharthi, B. Elizalde, H. Gamper, M. Al Ismail, R. Singh, B. Raj, and H. Wang, "PAM: Prompting audio-language models for audio quality assessment," in *Interspeech 2024*, 2024, pp. 3320–3324.

[12] H.-H. Wu, O. Nieto, J. P. Bello, and J. Salamon, "Audio-text models do not yet leverage natural language," in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.

[13] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[14] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: text-to-audio generation with latent diffusion models," in *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[15] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[16] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3590–3598.

[17] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 564–572.

[18] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh, "Toward verifiable and reproducible human evaluation for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 277–14 286.

[19] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *The annals of mathematical statistics*, pp. 50–60, 1947.

[20] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[21] R. G. Steel, "A multiple comparison rank sum test: treatments versus control," *Biometrics*, pp. 560–572, 1959.

[22] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 143–146.

[23] R. Flesch, "A new readability yardstick." *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.

[24] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "BYOL for Audio: Exploring pre-trained general-purpose audio representations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 137–151, 2022.

[25] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *Proc. ICASSP*. IEEE, 2022, pp. 896–900.

[26] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," in *International conference on artificial neural networks*. Springer, 2005, pp. 799–804.

[27] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "MBNet: MOS prediction for synthesized speech with mean-bias network," in *Proc. ICASSP*, 2021, pp. 391–395.

[28] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015.

[30] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. ICASSP*, 2023, pp. 1–5.

[31] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," in *Proc. ICASSP*, 2023, pp. 1–5.