Correlation of Fréchet Audio Distance With Human Perception of Environmental Audio Is Embedding Dependent

Modan Tailleur, Junwon Lee, Mathieu Lagrange, Keunwoo Choi, Laurie M Heller, Keisuke Imoto, Yuki Okamoto

Correlation of Fréchet Audio Distance With Human Perception of Environmental Audio Is Embedding Dependent

1st Modan Tailleur
École Centrale Nantes
CNRS, LS2N, UMR 6004
F-44000 Nantes, France
modan.tailleur@ls2n.fr

1st Junwon Lee
Gaudio Lab, Inc. / KAIST
Seoul / Daejeon, South Korea
junwon.lee@gaudiolab.com

2nd Mathieu Lagrange
École Centrale Nantes
CNRS, LS2N, UMR 6004
F-44000 Nantes, France
mathieu.lagrange@ls2n.fr

3rd Keunwoo Choi
Gaudio Lab, Inc.
Seoul, South Korea
keunwoo@gaudiolab.com

4th Laurie M. Heller
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA, U.S.
hellerl@andrew.cmu.edu

5th Keisuke Imoto
Doshisha University
Kyoto, Japan
keisuke.imoto@ieee.org

6th Yuki Okamoto
Ritsumeikan University
Kusatsu, Japan
y-okamoto@ieee.org

***Abstract*—This paper explores whether considering alternative domain-specific embeddings to calculate the Fréchet Audio Distance (FAD) metric can help the FAD to correlate better with perceptual ratings of environmental sounds. We used embeddings from VGGish, PANNs, MS-CLAP, L-CLAP, and MERT, which are tailored for either music or environmental sound evaluation. The FAD scores were calculated for sounds from the DCASE 2023 Task 7 dataset. Using perceptual data from the same task, we find that PANNs-WGM-LogMel produces the best correlation between FAD scores and perceptual ratings of both audio quality and perceived fit with a Spearman correlation higher than 0.5. We also find that music-specific embeddings resulted in significantly lower results. Interestingly, VGGish, the embedding used for the original Fréchet calculation, yielded a correlation below 0.1. These results underscore the critical importance of the choice of embedding for the FAD metric design.**

***Index Terms*—Environmental Sound Synthesis, Objective Audio Quality, Neural Audio Embeddings, Evaluation Metrics**

I. Introduction

Generative audio synthesis has become a popular research topic in which deep neural nets are typically driven by textual prompts [1]–[5]. Those systems must be evaluated on high-level perceptual features such as audio quality and alignment with categories for meaningful comparisons. However, evaluating synthetic audio through perceptual evaluation remains a cumbersome process, despite its validity.

To address this challenge, various metrics have been developed for use in prototyping and large-scale quality assessment [2], [6], [7]. Among these metrics, the Fréchet Audio Distance (FAD) [8] is widely used. FAD compares the distribution of a reference set with that of synthetic audio using VGGish embeddings [9].

Recent work [10] demonstrates that considering alternative embeddings that have been trained on music data is beneficial for assessing the quality of music generation systems.

Similarly, in this paper, we investigate whether changing the embeddings can lead to an increased correlation of the FAD with human perceptual ratings of both audio quality and perceived fit to categories of environmental sounds (i.e., general audio excluding speech and music). Given the fact that VGGish embeddings have been trained on environmental audio, we expected that VGGish embeddings would perform well, but our findings show the opposite. VGGish embeddings report low correlations, as do the embeddings trained on music data. Fortunately, we find that more recent embeddings specifically trained on environmental audio are quite effective, suggesting that the choice of the embedding is a crucial part of FAD metric design.

The rest of the paper is organized as follows. Section II delves into related works, followed by the presentation of the selected embeddings in Section III. Section IV presents the experiments and Section V the correlation results between the FAD scores when considering several state-of-the-art neural audio embeddings and perceptual evaluations using the DCASE Task 7 2023 dataset [11]. The code corresponding to this study is made publicly available.[1]

II. Related work

The Fréchet Audio Distance (FAD) [8] has been proposed as an adaptation of the Fréchet Inception Distance (FID) [12] for audio quality assessment. FID and FAD compare the distribution of two datasets in a given embedding space. VGGish [13] was originally proposed as the feature extractor for FAD. To evaluate a synthesis model, a set of desired audio serves as the reference. Firstly, two multivariate Gaussian distributions, which have the same means and covariances as the embedding sets, are considered. Then, the Fréchet distance between the two distributions r and t is calculated as follows:

$$FAD(r,t) = \|\mu_r - \mu_t\|_2 + \text{tr}\left(\Sigma_r + \Sigma_t - 2\sqrt{\Sigma_r\Sigma_t}\right) \quad (1)$$

[1]Code repository: https://github.com/mathieulagrange/dcaseFadEmbedding

where μ_x and Σ_x are respectively the mean and covariance matrix of a given distribution x. The FAD calculation compares the two datasets in terms of fit to domain with the comparison of means, but also in terms of diversity by including a form of covariance comparison in the equation. A low FAD score thus indicates that the two datasets contain similar sound sources, and a similar diversity. If the reference dataset can be considered of high audio quality, it is generally assumed that a low FAD distance implies that the evaluated dataset is also of good audio quality.

Whether it be for audio or image evaluation, using Fréchet distance has a few drawbacks. Trying to match an embedding space can lead to models of very different quality having similar Fréchet distance scores, as one could be fitting the embedding space more accurately without improving the perceptual quality [14]. The majority of embeddings used are trained representations, implying a potentially strong dependency on the dataset and task used for training. Consequently, for accurate evaluation using Fréchet Distance, the dataset under assessment must exhibit similarity to the training set of the embedding. This emphasizes the potentially significant influence of the embedding choice on the Fréchet Distance calculation process.

For image generation evaluation, Kynkäänniemi, Tuomas, et al. [14] showed that simply matching the top-N classifications histogram between the reference and generated set improved the FID score without further improvement in the generative model. This indicates the high dependency of FID on ImageNet [15], which is used to train the Inception embedding. The high dependency on the choice of embedding may be much worse in audio domain. ImageNet dataset has carefully tailored 1k classes to cover the wide range of in-the-wild images. Conversely, VGGish was trained to classify only 3k labels, that are not even necessarily relevant to sound. This may limit the generalizability of audio embeddings, depending on the training data size and task.

Gui et al. [10] explored the limitations of FAD in music generation, as VGGish-based FAD struggles to accurately predict the perceptual features of generated musical audio. The authors investigated different embeddings and found that VGGish yields notably poor FAD scores compared to alternative representations. Although they have shown that embeddings such as CLAP [6], [16] are more suitable for music generation, their effectiveness in improving the FAD metric for Environmental audio generation remains to be evaluated.

III. Embeddings

Experimental details on all embeddings examined are presented in Table I. Our objective with this selection is to investigate whether domain-specific embeddings significantly influence the relevance of the Fréchet Audio Distance (FAD) metric. We consider VGGish [17] as our baseline, as it has a proven record of use for FAD calculation. Next, we consider MERT [18] as a recent embedding primarily trained on music data, as well as a CLAP model trained specifically on voice and music. Given that those latter embeddings are not trained on environmental audio, we hypothesize that they should perform poorly. Other CLAP models and PANNs [19] models, on the other hand, are trained using environmental audio. As CLAP models are trained using partly PANN architectures, or at least have considered PANNs in their framework in their evaluation protocol, they are expected to outperform PANNs.

TABLE I
Description of Embedding Models. The size of the Receptive Field (RF) is the maximal duration of audio considered by the model to compute the embedding.

	Model Size	**Audio Rate**	**Embedding Size / Rate**	**RF Size**
VGGish [17]	72M	16 kHz	128 / 1 Hz	1 s
MERT [18]	72M	24 kHz	768 / 76 Hz	5 s
MS-CLAP [20]	158M	44 kHz	1024 / 1 Hz	7 s
L-CLAP [16]	158M	48 kHz	512 / 1 Hz	10 s
PANNs-CNN14-16k [19]	80M	16 kHz	2048 / .1 Hz	10 s
PANNs-CNN14-32k [19]	80M	32 kHz	2048 / .1 Hz	10 s
PANNs-WGM-Logmel [19]	80M	32 kHz	2048 / .1 Hz	10 s

A. VGGish

VGGish [17] is an audio classifier trained on a subset of a large audio dataset extracted from YouTube videos called YouTube-100M, which contains 350,000h of audio data with video-level class labels. YouTube-100M covers a wide range of audio content, spanning everyday sounds, sound effects, and music, captured in diverse real-world scenarios. It's worth noting that within this dataset the video-level classes may not necessarily be directly related to the audio content, as a source can be present in a video without generating any sound. VGGish uses log-mel spectrograms with 64 frequency bins and 10-ms hops as input, and it has about 70M parameters. Following the methodology outlined by Gui et al. [10], we employ the VGGish model to process 1-second audio segments with 50% overlap.

B. MERT

The Music undERstanding model with large-scale self-supervised Training (MERT) [18] generates embeddings learned with teacher-student methods, using a combination of teachers including an acoustic teacher based on Residual Vector Quantization - Variational AutoEncoder (RVQ-VAE) and a musical teacher based on the Constant-Q Transform (CQT). The student model is a BERT-style transformer encoder. The MERT models are specialized in music, being trained on an in-house private music dataset comprising 160k hours of audio data. Among the several available models, we chose the one with 95M parameters (MERT-95M).

C. PANNs

PANNs [19] are classifiers trained on AudioSet [21]. The dataset originated from YouTube videos and is around 5,000h long, with 527 different general sound classes. Unlike YouTube-100M, AudioSet employs automatic labeling with human verification at the audio level. The majority of these models leverage log-Mel spectrograms featuring 64 mel bins and 10-ms hops as input. Released in various sizes and trained at different sample rates, these models offer versatility in application. Among them, the CNN14 model emerges as the

most commonly utilized variant, with 16kHz (PANN-CNN14-16kHz) or 32kHz (PANN-CNN14-32kHz) sampling rates. However, according to their paper, the model that leads to the best accuracy is the Wavegram-Logmel-CNN (PANN-WGM-LogMel). This model uses the audio waveform as input, which is transformed into a learned spectro-temporal representation along with a Mel-spectrogram. Subsequently, both representations are fed into the rest of the network. PANN-CNN14 and PANN-WGM-LogMel models contain 80M parameters.

D. MS-CLAP

The embedding Contrastive Language-Audio Pretraining (CLAP) [6] is trained to learn multimodal representations, using both an audio and a text encoder. Symmetric cross-entropy loss was exploited for language and audio cross-modal contrastive learning. It uses PANN-CNN14-32kHz for audio encoding and BERT for text encoding. It is trained on 128,000 audio/caption of FSD50k, ClothoV2, AudioCaps and MACS, representing about 250h of audio. For this study, we use the CLAP model released in September 2023 called "sep 23".

E. L-CLAP

LAION has also trained a CLAP architecture [16] similar to the one used for MS-CLAP. They trained their best model using HTS-AT [22] as the audio encoder and RoBERTa [23] as the text encoder. They proved that training their model on LAION-Audio-630k and AudioSet with keyword-to-caption augmentation significantly improves the performances of their CLAP model. LAION-Audio-630k comprises a diverse range of 4,000 hours of audio recordings depicting human activities, natural sounds, and audio effects, sourced from eight publicly available websites. In total, their model for environmental audio called "630k-audioset-best" (L-CLAP-audio) is reported to be trained on about 10,000h of audio. They also released models specialized in music, which are trained on music and speech from their data collection. We chose the model "music audioset epoch 15 esc 90.14" (L-CLAP-mus) for comparison.

IV. Experiments

A. Data

In this study, we leverage the DCASE 2023 Challenge Task 7 dataset [11], encompassing 700 sound excerpts of 7 different categories: *dog bark, footstep, gunshot, keyboard, moving motor vehicle, rain, and sneeze/cough*. Each sound excerpt is a mono 16-bit 22,050 Hz 4-second audio sourced from three distinct datasets: UrbanSound8K [24], FSD50K [25], and BBC Sound Effects[2]. To ensure high relevance, diversity, and clarity, the challenge organizers manually selected and validated the excerpts.

Supplementary to this dataset are the audio generated using the baseline and 8 submitted algorithms, with each contributing an additional set of 700 sound excerpts synthesized by their respective systems. The duration of the whole dataset (recorded and synthesized) is about 8h. The 8 systems from the participants were top-ranked in terms of FAD score. Twenty sounds from each system underwent perceptual evaluation for both audio quality and category fit by 91 raters (47 hrs). Each category within each system has been evaluated, resulting in a total of 63 evaluations for each criterion. The correlations shown in section V are thus based on those 20 evaluated audios per category and per system.

[2]https://sound-effects.bbcrewind.co.uk/

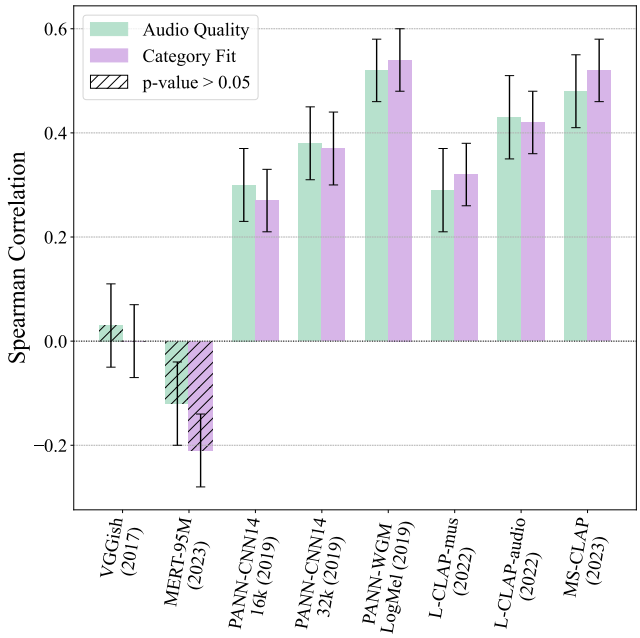

Fig. 1. Spearman correlation coefficient ($n = 63$) between FAD^{-1}. and perceptual evaluation of audio quality and category fit for different embeddings. Error bars display standard deviation.

B. Uncertainty estimation

To assess the uncertainty associated with each Spearman correlation calculation, we introduce Gaussian noise with a standard deviation of 1 to the perceptual evaluation scores. Subsequently, we repeat this process 100 times to generate 100 different noisy sets, each containing 63 perceptual evaluations for both category fit and audio quality.

By computing the mean and standard deviation of the Spearman correlation coefficients across these 100 noisy sets, we obtain estimates of the variability and uncertainty inherent in the correlation calculations.

V. Results

Given the FAD definition given in Eq. 1, lower FAD scores indicate a high similarity between datasets. Consequently, a reliable FAD metric should demonstrate an inverse correlation with high perceptual quality when computed between a reference set and various generated sets. To simplify the presentation of results, we use the inverse of the FAD (FAD^{-1}) so that a high quality FAD^{-1} metric should achieve a high positive correlation with perceptual attributes that are higher when better.

A. Overall Correlation

As shown in Figure 1, the PANNs-WGM-LogMel FAD and the MS-CLAP FAD demonstrate strong correlations with both category fit and audio quality, with PANN-WGM-LogMel

being significantly higher than MS-CLAP FAD. In contrast, both the MERT-95M FAD and the VGGish FAD demonstrate very weak correlation with perceptual evaluation. Additionally, the L-CLAP models perform less effectively than MS-CLAP and PANN-WGM-LogMel, with L-CLAP-audio showing a better correlation score than L-CLAP-mus.

B. Per-category Correlation

In Figure 2, we present the per-category correlation results. Interestingly, substantial variability in correlation is observed across different categories. PANNs-WGM-LogMel displays greater stability across categories, while performing better than the other embeddings in half of the categories. VGGish demonstrates good performance in categories such as sneeze/cough and gunshot, but performs extremely poorly in every other category. Similarly, CLAP exhibits low correlations in categories such as moving motor vehicle, sneeze/cough, and gunshot, while demonstrating better performance in other categories, competing or beating PANNs-WGM-LogMel in some of them. Overall, MS-CLAP and PANN-WGM-LogMel lead to higher correlations than VGGish across nearly all categories. These results should be treated cautiously because the Spearman correlation coefficients are calculated with only 9 data points per category.

C. Influence of Dimensionality

As shown in Table I, each embedding model varies in size. Considering the findings presented in Figure 1, we want to investigate if the dimensionality of the embedding may bias the performance of the FAD metric. To refute the hypothesis that a higher dimensional embedding may gain an unfair advantage, we conduct a dimensionality reduction of every embedding to match that of VGGish (n=128). The dimensionality reduction is performed by projecting the set of embeddings on the 128 Eigenvectors of the highest Eigenvalues using Principal Component Analysis (PCA).

We observed that diminishing the size of the embeddings had minimal impact on the correlations with FAD^{-1}, decreasing them by approximately 0.01 for each embedding. Thus, the size of the embedding does not significantly influence the performance of the FAD metric.

D. FAD-based Category Mapping

Little is known about how the FAD may relate to the similarity of sound categories. To examine this, we grouped the 7 sound categories of the 700 audios of the evaluation set of DCASE Task 7 2023 dataset into 3 meta-categories: Impact (Footsteps, Gunshot, Keyboard), Vocalizations (Dog Bark, Sneeze/Cough), and Texture (Moving Motor Vehicle, Rain). A satisfactory embedding should result in the FAD grouping similar categories together while separating highly dissimilar categories. Given that the 7 sound classes themselves are highly distinct, they should also be quite separated from each other in the projection.

Figure 3 shows a 2D mapping of the similarities of every pair of categories from the DCASE Task 7 2023 dataset, calculated with Multidimensional Scaling (MDS) with FAD using three different embeddings as input. We find that the three

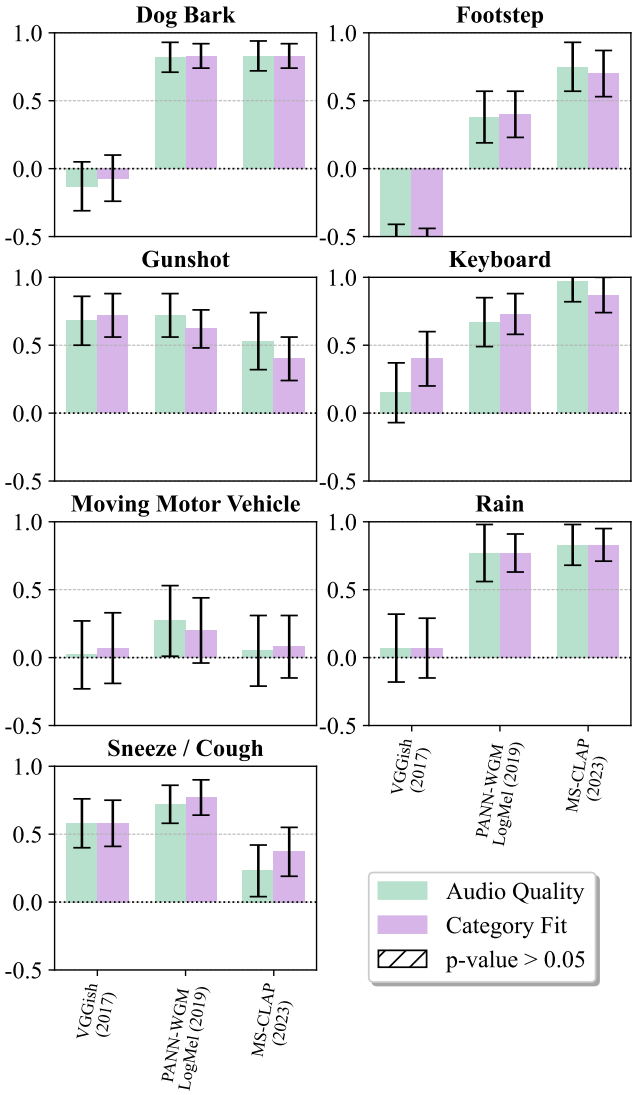

Fig. 2. Spearman correlation coefficient ($n = 9$) between FAD^{-1}. and perceptual evaluation of audio quality and category fit for VGGish, PANNs CNN14 Wavegram Logmel and CLAP.

embeddings successfully group similar categories together, as evidenced by the Voronoi separation diagram. Additionally, the split between the different groups is somewhat more pronounced for MS-CLAP and PANN-WGM-LogMel than for VGGish. Furthermore, for MS-CLAP and PANN-WGM-LogMel, all 7 sound sources are further apart from each other, which indicates that these two embeddings are better at separating the different categories than VGG.

VI. Conclusion

In this paper, we explored the use of alternative embeddings to assess audio quality and alignment with categories in environmental audio, aiming to improve the validity of the Fréchet Audio Distance (FAD) metric. We compared several embeddings, including VGGish, PANNs, MS-CLAP, L-CLAP, and MERT, using the DCASE Task 7 2023 dataset.

We find that there is a strong dependency of the embedding in the FAD metric, which appears to be closely linked to the domain of the embedding training dataset. In fact, music-

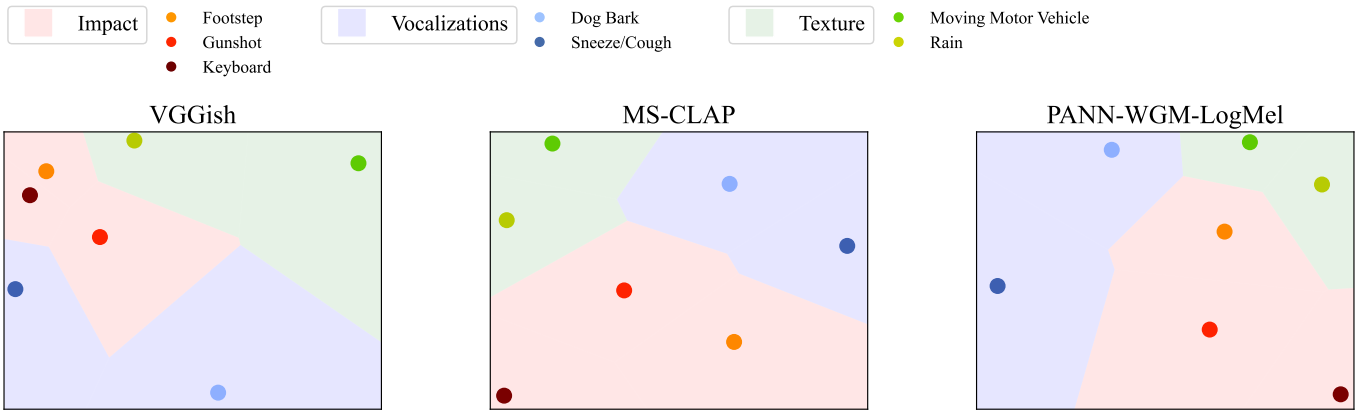

Fig. 3. 2D Projection of inter-category FAD similarity matrix using Multi-dimensional scaling (MDS) on DCASE Task 7 2023 dataset.

trained embeddings, such as MERT-95M and CLAP Laion Music, perform less effectively than those trained on environmental audio. Furthermore, while VGGish was trained to classify 3k video-level labels, these labels might not necessarily be relevant to sound, potentially limiting its generalizability for tasks like DCASE TASK 7 2023 systems evaluation. This observation also aligns with findings reported by Gui et al. [10].

CLAP and PANN-WGM-LogMel clearly outperform VGG in correlating with the perception of environmental audio. Although there is variation in performance across categories, overall the PANN-WGM-Logmel slightly but significantly outperforms CLAP. However, because CLAP outperforms PANNs in sound event classification tasks when both use the PANN-CNN14-32kHz model, [6] we believe future CLAP models trained with the more advanced PANN-WGM-Logmel model [19] may show superior performance.

We investigated the influence of deep audio embeddings in the formation of a metric space that reflects the high-level organization of sound events. The preliminary experiments presented here demonstrate the advantages of recent embeddings and specialized embeddings tailored to specific tasks. Further investigation is recommended, for example, by considering a more diverse number of categories for which perceptual ratings are available.

References

[1] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, 2016.

[2] 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," *arXiv preprint arXiv:2301.12503*, 2023.

[3] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," in *ICLR*, 2023.

[4] 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.

[5] 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 *ICML*, 2023.

[6] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: learning audio concepts from natural language supervision," in *ICASSP*, 2023.

[7] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, "Mulan: A joint embedding of music audio and natural language," in *ISMIR*, 2022.

[8] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms," in *Interspeech*, 2019.

[9] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "CNN architectures for large-scale audio classification," in *(ICASSP)*, 2017.

[10] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting frechet audio distance for generative music evaluation," in *ICASSP*, 2024.

[11] K. Choi, J. Im, L. M. Heller, B. McFee, K. Imoto, Y. Okamoto, M. Lagrange, and S. Takamichi, "Foley sound synthesis at the DCASE 2023 challenge," in *DCASE Workshop*, 2023.

[12] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," *NeurIPS*, 2017.

[13] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, "CNN architectures for large-scale audio classification," in *(ICASSP*, 2017.

[14] T. Kynkäänniemi, T. Karras, M. Aittala, T. Aila, and J. Lehtinen, "The role of imagenet classes in fréchet inception distance," in *ICLR*, 2023.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2009.

[16] 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 *(ICASSP)*, 2023.

[17] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, and B. Seybold, "CNN architectures for large-scale audio classification," in *(ICASSP)*, 2017.

[18] Y. LI, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Z. Wang, Y. Guo, and J. Fu, "MERT: Acoustic music understanding model with large-scale self-supervised training," in *ICLR*, 2024.

[19] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, 2020.

[20] B. Elizalde, S. Deshmukh, and H. Wang, "Natural Language Supervision for General-Purpose Audio Representations," Sept. 2023. arXiv:2309.05767 [cs, eess].

[21] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *(ICASSP)*, 2017.

[22] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection," in *(ICASSP)*, 2022.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[24] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014.

[25] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50K: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, 2021.