Effective Techniques for Scaling Audio Encoder Pretraining

Byeonggeun Kim*[1], Andrew Bydlon*[2], Qingming Tang[2], Huy Phan[3], Chieh-Chi Kao[1], Tao Zhang[2], Chao Wang[2]
Amazon AGI
[1]Bellevue, United States; [2]Boston, United States; [3]Cambridge, United Kingdom
{bgkim, abydlon, qmtang, huypq, chiehchi, taozhng, wngcha}@amazon.com

***Abstract*—This work presents advancements in audio pretraining objectives designed to generate semantically rich embeddings, capable of addressing a wide range of audio-related tasks. Despite significant progress in the field, current methods often emphasize full fine-tuning in downstream applications, which can obscure the true potential of pretrained audio encoders. In this study, we present an audio encoder that achieves state-of-the-art (SOTA) performances in both fine-tuning and linear probing, utilizing a carefully curated set of pragmatic techniques. Building on previous research, we incorporate masked prediction and introduce SpecAug within a curriculum masking strategy at the patch level, which progressively increases training difficulty, along with a mask-aware position bias. To comprehensively assess the encoder's capabilities, we examine the impact of scaling both the dataset size and model capacity, conducting linear probing evaluations while keeping the encoder frozen as well as full fine-tuning. Our model demonstrates superior performance compared to recent SOTA methods across various downstream tasks. Additionally, we explore the potential of tokenizing the resulting audio embeddings for use as discrete inputs, enhancing our understanding of the model's capabilities.**

***Index Terms*—Audio Pretraining, Audio Encoder, Semantic Embedding, Tokenization, Masked Prediction**

I. Introduction

Audio pretraining aims to develop a comprehensive embedding space, producing semantically rich embeddings that support a wide range of downstream applications [1], [2]. Recent advancements in audio pretraining have achieved significant progress leveraging large-scale data, and resulting in more generalizable embeddings. Building on this foundation, self-supervised learning such as contrastive learning with multi-modal alignment [3], [4] or masked prediction [5]–[7], as well as weakly supervised learning [8] techniques have been applied for audio encoder pretraining.

In the realm of weakly supervised learning, Whisper [8] demonstrated the efficacy of scaling data by utilizing extensive transcripts of audio on the internet. On the other hand, AudioMAE [5] proved that applying masked part prediction using the unmasked section, with a high masking ratio, is also effective in audio domain. Building on the masked prediction approach, BEATs [6] introduced an iterative training scheme that begins by generating target tokens with a random tokenizer and later employs a self-supervised trained tokenizer. The training objective involves both mask and discrete label prediction, leading to promising performance but at the cost of increased computational intensity due to its iterative training process. A-JEPA [7] introduced a curriculum masking strategy using a context encoder to predict patch-wise representations of regions sampled at specific locations, achieving performance comparable to that of BEATs [6]. Recently, another line of research has benefited from multimodal alignments through contrastive learning between encoders of other modalities [3], [4]. Leveraging multimodal alignments has been shown to enhance the understanding within each modality [9], [10]. However, this approach represents a parallel effort to unimodal training and is outside the scope of this work. Our method, through focused on unimodal training, could potentially be synergized with multimodal approaces to further enhance peformance.

Despite of the promising application of recent audio encoders in various downstream tasks [1], [2], current approaches often fall short in fully understanding the capabilities of the resulting embeddings. This limitation arises because evaluations are typically based on full-finetuning tasks using predefined, limited training datasets and model sizes [5]–[7]. Full fine-tuning results can be sensitive to training configurations, which may obscure the true capabilities of the pretrained model. To address this gap, we investigate the potential of an audio encoder through a pretraining approach that achieves state-of-the-art (SOTA) performance in both fine-tuning and linear probing tasks, employing a carefully selected set of techniques. Building on patch-wise representations with a masked prediction objective, our training method begins by augmenting an input audio sample into two randomly augmented versions and then introduces a masking strategy within a curriculum that gradually increases the training objective's difficulty. This strategy combines structured and random masking, implementing bandwise masking inspired by SpecAug [11] at the patch level. Additionally, we incorporate a mask-aware position bias to further enhance the training process. We also explore the impact of scaling both model capacity and data. As a result, our model not only achieves SOTA performance in full fine-tuning but also excels in linear probing, maintaining the audio encoder in a frozen state across various tasks. Furthermore, we analyze its effectiveness as a source of discrete tokens by tokenizing the resulting embeddings, allowing us to gain a deeper understanding of its capabilities.

* equal contribution

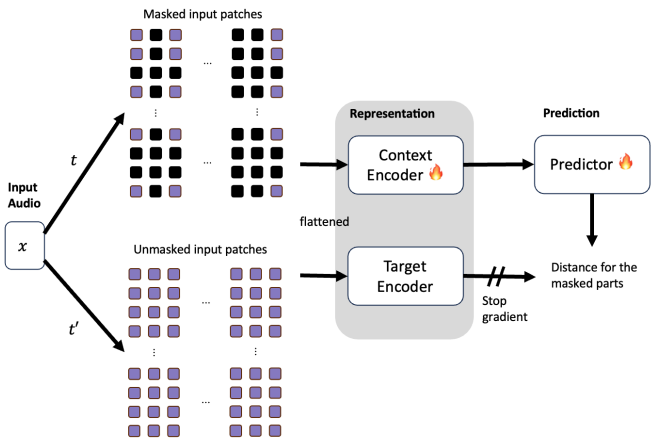

Fig. 1. **Training process of the audio context encoder.** The inputs to the context and target encoders are randomly augmented, followed by patch-level masking for the context encoder and no masking for the target encoder. The training objective is to predict the subset of the target encoder's embeddings corresponding to the masked portions.

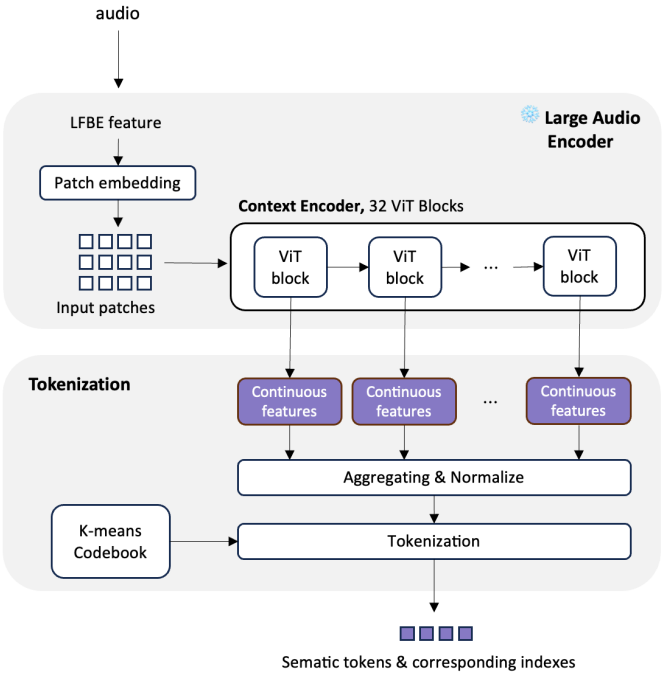

Fig. 2. **Tokenization Process.** The representation across the layers of the audio encoder are aggregated and normalized. The resulting embedding is tokenized using the K-means codebooks.

II. Methods

A. Audio Encoder Pretraining

The Figure 1 illustrates the training process of the audio context encoder. Given a distribution of audio augmentations $\mathcal{T}$, we begin by applying transformations $t \sim \mathcal{T}$ and $t' \sim \mathcal{T}$ to the input audio x, resulting in the augmented views $t(x)$ and $t'(x)$, respectively.

Input Patch-level Representation. Our audio encoder training method incorporates mask prediction techniques as described in [6], [7]. Starting with raw augmented audio inputs, $t(x)$ or $t'(x)$, we extract LFBE features. Following the framework of Vision Transformers (ViT) [12], LFBE features are segmented into non-overlapping patches and linearly projected into a D-Dimensional embedding space. For instance, a 9.94 second audio clip sampled at 16 kHz is converted into 128-dimensional LFBE features using a 25ms window and a 10ms hop length, resulting in 992×128 outputs ($T \times F$). From these LFBE features, we apply a non-overlapping 16×16 convolutional layer with a stride of 16 to obtain 62×8 patches. These patches are then arranged in time-major flattening, which are subsequently processed by an encoder. We denote the resulting D-dimensional patches of an augmented view, $t(x)$ as $V = \{v_i\}_{i=1}^{L}$, where $L = T \times F$, and T and F represent the dimensions along the time and frequency axes, respectively. Similarly, we denote the patches for $t'(x)$ as V'.

Curriculum Masking. Drawing inspiration from [7], [11], we implement a curriculum masking strategy on patches V. Our strategy progressively increases in difficulty as training progresses, achieved through two methods: (1) elevating the mask ratio and (2) intensifying the task complexity. Initially, we set the masking rate at 0.65 gradually advancing it to 0.80 by the end of training. Additionally, we employ two types of masking techniques for the task complexity: structured and random masking. Structured masking involves masking entire time and frequency-wise patches, inspired by [11], while random masking entails random patch masking. Throughout training, we augment the utilization of structured masking compared to random masking, escalating it from 0.05 to 0.75. We observed that as structured masking becomes more prevalent during training, it leads to increased training loss (reflecting a more challenging task compared to random masking) and results in superior generalization and performance. Following, MAE [13], the context encoder, f_θ, only processes the unmasked parts of V, denoted as $\mathcal{M}(V)$.

Mask-Aware Position Bias. While applying the convolution-based relative position embedding layer at the bottom [6] and the gated relative position bias [31], we adapt the position bias to be aware of the curriculum masking. This is achieved by first calculating the position bias for the full patch length, L, and then applying the same mask to the position bias. Our empirical findings indicate that this mask awareness is a crucial component for achieving optimal performance.

Training Objective. Leveraging learning schemes with masked prediction [6], [7], [13], we process the masked input $\mathcal{M}(V)$ and the unmasked input V' using the context encoder, f_θ, and the target encoder, $f_{\theta'}$, respectively. The θ' is updated using the exponential moving average (EMA) method [20], [21], based on the θ. Our objective is to predict the subset of $f_{\theta'}(V)$ corresponding to the masked sections of V using a predictor g_ϕ. The training objective is then defined as the conventional mean squared error (MSE) loss between the prediction, $g_\phi(f_\theta(\mathcal{M}(V)))$, and its target $f_{\theta'}(V)$,

TABLE I
COMPARATIVE ASSESSMENT. MEAN AVERAGE PRECISION (MAP: %) FOR AUDIOSET (AS) [14] AND CLASSIFICATION ACCURACY (%) FOR OTHER FIVE DOWNSTREAM TASKS [15]–[18].

Method	#Param	Training Data	Full-Finetune	Linear Probing				
			AS-2M [14]	AS-20K [14]	GTZAN [15]	CochlScene [16]	Meld [17]	VocalSound [18]
BEATs, iter 3+ [6]	90M	AudioSet [14]	48.6	-	-	-	-	-
A-JEPA [7]	86M	AudioSet [14]	48.6	-	-	-	-	-
Reproduce BEATs [6]	90M	AudioSet [14]	48.4	26.2	76.0	79.6	49.8	85.5
Reproduce BEATs [6]	435M	AudioSet [14]	49.1	32.4	79.9	81.7	54.8	88.2
MERT [19]	330M	MusicData	-	-	79.3	-	-	-
Whisper Large V2 [8]	1.55B	Multi-Datasets	45.7	32.8	-	-	-	-
Ours-Small	90M	AudioSet [14]	48.1	30.8	78.5	81.4	55.4	90.8
Ours-Medium	411M	Multi-Datasets	**49.2**	**33.9**	84.4	**84.2**	57.2	90.1
Ours-Large	1.1B	Multi-Datasets	48.4	32.5	**85.7**	83.8	**57.4**	**90.8**
Qwen-Audio [1]	8.55B	Multi-Datasets	-	-	-	79.5	55.7	92.9

formulated as:

$$\theta^*, \phi^* = \underset{\theta,\phi}{\operatorname{argmin}} \{\mathbb{E}_x \mathcal{L}_{\text{MSE}}(g_\phi(f_\theta(\mathcal{M}(V))), f_{\theta'}(V'))\}. \quad (1)$$

Please note that we concatenate zero vectors for the masked parts of $f_\theta(\mathcal{M}(V))$, as described in [13], to serve as input to g_ϕ. After training, the resulting context encoder f_θ is used as the final audio encoder.

B. Tokenization

After pretraining, we experimented with tokenizing the resulting audio embeddings to explore their potential use as discrete tokens in downstream tasks [22]. We applied K-means clustering, a widely used method for tokenization audio embeddings [22], [23]. Instead of relying solely on the embeddings from the last layer, we utilized a simple averaging of embeddings from several of the final layers. We performed clustering on the training dataset and used the resulting centroids for tokenization. The overall process is depicted in Fig 2. This approach enabled us to assess the impact of tokenization on audio understanding tasks.

III. EXPERIMENTS

A. Experimental Setup

Scaling Data & Model. While recent audio encoder approaches have typically been trained and evaluated using AudioSet's [14] 5.6K hours of data, we explored the effects of scaling the dataset size by a factor of 70, resulting in a total of 400K hours. Due to the limited size of publicly available datasets [14], [24]–[26], we expanded our dataset by incorporating proprietary data. Table II provides a summary of the data in our experiments. Additionally, we investigated model scaling by examining models with three different parameter scales: 90M, 411M, and 1.1B. The scaling was conducted both in terms of width and depth (from 90M to 411M) and solely in terms of width (from 411M to 1.1B).

Model Architecture. The context and target encoders share same architecture, based on the ViT [12]. Specifically, our 411M model comprises 32 ViT blocks, with 1,024-dimensional input patches, a 1,024-dimensional hidden layer, a 4,096-dimensional feed-forward network, and 16 attention heads using GELU [27] activation. We did not use layer-wise gradient decay or dropout during pretraining. The predictor g_ϕ utilizes the same architecture but with fewer blocks, 4.

TABLE II
OVERVIEW OF EXPERIMENTAL DATASET SCALES

Dataset	# Samples	# Hours	Dataset	# Samples	# Hours
AudioSet [14]	2.03 M	5.6 K	CHIME-6 [26]	5e-4 M	1.4 K
VGG-Sound [24]	0.20 M	0.5 K	MusicBench [25]	0.05 M	0.1 K
Internal Data	8.92 M	392.4 K			
Total	11.2 M	400 K			

Training Setup. We trained the model using 9.935-second audio segments sampled at 16kHz, for a total of 430K steps with a batch size of 20.5k seconds for pretraining, and 200K steps with a batch size of 5.1k seconds of audio for full fine-tuning. The training was conducted with binary float 16 (BF16) precision. We used the AdamW optimizer [28] with a linear decay scheduler, starting with a warm-up phase for 10 % of the total steps, and peaking at learning rates of 1e-4 for pretraining and 5e-5 for full fine-tuning, applying gradient clipping at a value of 2. Data augmentation was performed using a combination of techniques, including circular time shift, energy shift, and various audio effects such as high-pass, low-pass, band-pass, and band-reject filters, as well as crystalizer, flanger, phaser, and Haas effects, collectively represented by $\mathcal{T}$. For full fine-tuning, we further utilized SpecAug [11].

Evaluation Metrics. To provide a more direct assessment, we compare our method to recent approaches by using linear probing using the weighted sum of all layers' outputs [29] while keeping the audio encoder frozen rather than full finetuning. We evaluate our method across five downstream tasks: sound event classification on AudioSet [14], music genre recognition on GTZAN [15], audio scene classification on CochlScene [16], speaker emotion recognition on Meld [17], and vocal sound classification on VocalSound [18]. We utilized a 1- or 2-layer MLP and reported the results for the configuration that yielded the best performance for each.

B. Experimental Results

Table I presents a comparative assessment of our audio encoder against other state-of-the-art models. BEATs [6], one

TABLE III
IMPACT OF TOKENIZATION ON OURS-MEDIUM (411M): MAP (%) ON AS [14] AND ACC. (%) ON OTHER DATASETS.

Method	bps	AS-20K	GTZAN	CochlScene	Meld	VocalSound
Continuous	-	33.9	84.4	84.2	57.2	89.1
k-means clustering	764	24.3	74.1	78.3	55.7	88.9
+ avg Freq	96	22.0	75.9	72.5	52.9	87.0
+ Audio type ensemble	333	22.9	68.8	75.5	54.8	88.4
+ Dual-tower tokens.	166	-	85.7	77.0	54.9	87.2

TABLE IV
ABLATIONS: MAP (%) ON AS [14] AND ACC. (%) ON OTHER DATASETS.

Pos	Mask	Scale	AS-20K	GTZAN	CochlScene	Meld	VocalSound
			25.3	73.2	78.9	55.3	85.8
✓			26.5	75.4	79.5	55.3	85.7
✓	✓		30.8	78.5	81.4	55.4	90.8
✓	✓	✓	33.9	84.4	84.2	57.2	89.1

of the latest SOTA audio encoders, was included in the comparison. Due to the lack of reported linear probing results for [6], we reproduced it, scaling it to 430M parameters as well.

First, we trained exclusively on AudioSet [14] with 90M parameters for a fair comparison to [30] and [7]. Our model surpassed the BEATs 90M across all five linear probing tasks by a clear margin, despite some degradation in full fine-tuning results, which tend to be sensitive to fine-tuning configurations. Next, our medium-sized 411M model, scaled in both model size and data, not only outperformed all other methods in full-finetuning results on AudioSet but also exceeded all baselines across five linear probing tasks. This demonstrates the superior quality of our pretrained embeddings.

Notably, our model was general enough to surpass music-specific encoders like MERT [19] in the Music Genre Classification task on GTZAN [15], with a significant margin (79.3 vs. 84.4). Additionally, our encoder outperformed the Whisper Large V2 [8], a strong baseline with more parameters (1.8B), in AudioSet tasks. We further scaled the model to 1.1B parameters, though the improvements over our 411M model were marginal.

Finally, we compared our results with Qwen-Audio [1], which leverages an 8B LLM for classification tasks. Our results indicate that the resulting embeddings from our model are strong enough to be comparable to those from an Audio Encoder + LLM setup, even with significantly fewer parameters.

C. Tokenization

Table III shows the performance after tokenization. We used K-means clustering approach with 40,960 centroids, resulting in 0.8 kbps (calculated as $\frac{1}{0.16} \times 8$ (freq dim) $\times \log_2 40960$). The linear probing results indicate a significant information loss after tokenization, particularly for AudioSet, where there was a 28.3 % relative performance drop. We further experimented with aggregating the frequency dimension, "avg Freq", to reduce the token/sec rate and lessen the burden on subsequent usages, but this approach led to additional performance degradation across most tasks.

To mitigate the performance drop, we first attempted an ensemble method by splitting the training data into categories—music, human speech, other sounds, and web-crawled datasets—using 10,240 centroids for each category [23]. The resulting tokens were concatenated, which led to a fourfold increase in token/sec. While this approach resulted in performance improvements for Meld, VocalSound, and AudioSet, it caused a performance drop in other tasks, indicating the need for careful tuning regarding the splitting strategy and the ratio of centroids.

Alternatively, we explored the use of a dual-tower token architecture on top of the "k-means clustering + avg Freq" approach. In this setup, one tower of tokens is derived from the pretrained model, while the other is generated by the AudioSet-2M fully finetuned model, with each tower using 10,240 centroids. This approach is particularly effective in the GTZAN and CochlScene benchmarks, demonstrating the benefits of combining two different types of tokens. This is similar to the approach in [23], where a fine-grained acoustic token is paired with a coarse semantic tokens.

D. Ablations

Table IV presents an ablation study that evaluates the effectiveness of the proposed components: (1) Pos: mask-aware position bias, (2) Mask: SpecAug-style curriculum masking, and (3) Scale: scaling of both the model (from 90M to 411M parameters) and the data. The results indicate that the masking-aware positional bias provides clear benefits across most tasks, except for Vocal Sound when random masking is applied. Additionally, the proposed curriculum masking demonstrates a significant advantage in most tasks, with the exceptions of Meld. Finally, the study shows that the proposed pretraining method is highly scalable in terms of model parameters and data, yielding clear benefits in most tasks except for Vocal Sound.

IV. CONCLUSION

In this study, we investigated audio encoder pretraining using a masked prediction scheme. The proposed training method begins by augmenting an input audio sample into two randomly augmented versions and performing masked prediction at the patch level. Instead of relying solely on random masking, we introduced a curriculum masking approach that combines SpecAug-style structured masking with random masking incorporating mask-aware position bias. This training scheme proved to be effective, enabling us to scale the model in terms of both size and data. As a result, the audio encoder achieved state-of-the-art (SOTA) performance in linear probing across various downstream tasks, demonstrating its robustness and generalization capabilities. Additionally, we investigated the tokenization of the resulting embeddings, which provided deeper insights into the model's capabilities.

REFERENCES

[1] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," CoRR, vol. abs/2311.07919, 2023.

[2] Z. Cheng, S. Leng, H. Zhang, Y. Xin, X. Li, G. Chen, Y. Zhu, W. Zhang, Z. Luo, D. Zhao, and L. Bing, "Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms," 2024.

[3] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "CLAP learning audio concepts from natural language supervision," in IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023.

[4] 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 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023.

[5] P.-Y Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, "Masked autoencoders that listen," in Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022.

[6] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "Beats: Audio pre-training with acoustic tokenizers," in International Conference on Machine Learning, ICML 2023.

[7] Z. Fei, M. Fan, and J. Huang, "A-JEPA: joint-embedding predictive architecture can listen," CoRR, vol. abs/2311.15830, 2023.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in International Conference on Machine Learning, ICML 2023.

[9] P.-Y Huang, V. Sharma, H. Xu, C. Ryali, H. Fan, Y. Li, S.-W. Li, G. Ghosh, J. Malik, and C. Feichtenhofer, "Mavil: Masked audio-video learners," in Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023.

[10] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, C. Zhang, Z. Li, W. Liu, and L. Yuan, "Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment," in The Twelfth International Conference on Learning Representations, ICLR 2024.

[11] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in Interspeech 2019.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in 9th International Conference on Learning Representations, ICLR 2021.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. B. Girshick, "Masked autoencoders are scalable vision learners," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022.

[14] J. F. Gemmeke, D. P. W. 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 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017.

[15] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," CoRR, vol. abs/1306.1461, 2013.

[16] I.-Y. Jeong and J. Park, "Cochlscene: Acquisition of acoustic scene data using crowdsourcing," CoRR, vol. abs/2211.02289, 2022.

[17] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019.

[18] Y. Gong, J. Yu, and J. R. Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022.

[19] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. B. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Y. Guo, and J. Fu, "MERT: Acoustic music understanding model with large-scale self-supervised training," CoRR, vol. abs/2306.00107, 2023.

[20] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Ávila Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - A new approach to self-supervised learning," in Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020.

[21] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020.

[22] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "Audiolm: A language modeling approach to audio generation," IEEE ACM Trans. Audio Speech Lang. Process., vol. 31, pp. 2523–2533, 2023.

[23] H. Liu, X. Xu, Y. Yuan, M. Wu, W. Wang, and M. D. Plumbley, "Semanticodec: An ultra low bitrate semantic audio codec for general sound," 2024.

[24] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020.

[25] J. Melechovsky, Z. Guo, D. Ghosal, N. Majumder, D. Herremans, and S. Poria, "Mustango: Toward controllable text-to-music generation," CoRR, vol. abs/2311.08355, 2023.

[26] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj, et al., "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in CHiME 2020-6th International Workshop on Speech Processing in Everyday Environments, 2020.

[27] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," arXiv preprint arXiv:1606.08415, 2016.

[28] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in 7th International Conference on Learning Representations, ICLR 2019.

[29] S.-W. Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H.-y. Lee, "SUPERB: speech processing universal performance benchmark," in 22nd Annual Conference of the International Speech Communication Association, Interspeech 2021.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019.

[31] Z. Chi, S. Huang, L. Dong, S. Ma, B. Zheng, S. Singhal, P. Bajaj, X. Song, X.-L. Mao, H. Huang, and F. Wei, "XLM-E: Cross-lingual Language Model Pre-training via ELECTRA". ACL (1) 2022: 6170-6182