From Logocentrism To Unification Modeling Of All Audio And Music Tasks

Xingjian Du duxingjian.real@bytedance.com

Abstract

Audio signal processing has entered the Deep Learning (DL) era. However, its relation with other DL fields, like computational vision and natural language processing are under investigated. In this proposal, inspired by the idea of logocentrism, I attempt to take advantage of large language models as the logos model to unify all audio and music tasks under a single paradigm. Specifically, according to the logocentrism, audio and music belong to the category of external reality, thus the audio and music tasks can be described and modelled in the form of language naturally. Besides, with the rapid development of NLP fields, there is evidence showing that large language models do have reasoning ability in some sense. Thus, I propose and implement a logocentrism-based model which can be transferred to multiple music tasks. Experimental results on two tasks demonstrate the feasibility of my proposal.

1 Introduction

Deep learning models have achieved great success in many audio and music-related research fields, such as sound event detection [1], source separation [2], genre classification [3], melody extraction [4] and version identification [5]. Currently, most approaches are task-specialized, which leads to dedicated architectures and isolated models.

The drawbacks of task specified solutions are obvious. First, it is not elementary to develop architectures for a large number of audio tasks under different settings. Second, learning isolated models severely restricts the knowledge sharing between related tasks and settings. Finally, it is costly and time-consuming to construct data sets and corpora specialized for different audio tasks. The idea of logocentrism motivated me to propose a task-agnostic paradigm to solve these problems with a unification framework for all audio and music subtasks. The logoscentrism refers to the tradition of Western science and philosophy that regards words and language as a fundamental expression of an external reality. It holds the logos as epistemologically superior and that there is an original, irreducible object which the logos represent [6]. The audio and music belong to the category of *external reality*, thus the audio and music tasks can be described and modelled in the form of language naturally. Meanwhile, the research on Large Language Models (LLMs) has made tremendous progress [7] in these five years. And the sign of reasoning ability emerges as the number of parameters in these LLMs grows dramatically. Thus, the implementation of a new learning paradigm motivated by logocentrism is feasible in engineering. Under this paradigm, the modelling of various audio and music sub-tasks can be decoupled into two parts: (1) an encoder module that transcribes the audio to the representation of logos, i.e. the embedding of

human languages. (2) the core reasoning module that performs reasoning on the mixture of native text inputs and quasi-text embedding transcribed by aforementioned encoder module.

2 Details of Logocentrism Unification Learning Framework

Following the idea of logocentrism unified learning framework introduced in the Sec.1, the abstract framework is built on two separate parts as the Fig.1 shows.

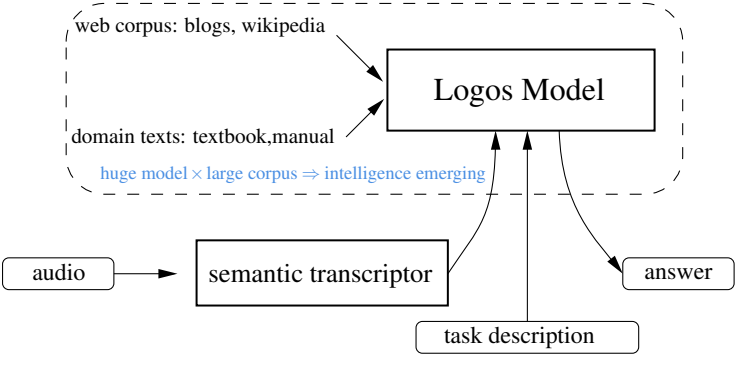


Figure 1: The illustration of logocentrism unification framework, which is comprised of a audio semantic transcriptor and logos model.

The significant difference with existing end-to-end learning manners is that the logocentrism framework adopts a decoupled pipeline. The core idea behind this disentangled paradigm comes from logocentrism, and it aims to formalize learning as the combination of the perception of external reality (audio specifically) and introspective reasoning on the mixture of language-related representation of audio and task-related description. The advantage of this design is that it provides a task description mechanism that makes one framework adopts all tasks while task-related texts can control the function of the unified model.

Moreover, the isolated development of logos reasoning models enables learning from pure human language or audio-text paired data. The former can be collected by web crawling and the parsing of domain text materials for the learning of common sense and domain knowledge jointly. The following parts of this section include a brief survey about other attempts at audio-text modelling and depiction of the proof of concept (PoC) implementation of the proposed framework.

2.1 Related Work

There are already some methods that introduce language models for better audio task resolving. However, recent work for audio-text translation is mainly focused on a single task, especially transcription[3, 9]. [8] utilizes a freeze GPT2 [7] as a deocder for few-shot learning ability. Such a specialized setting ensures the model's performance while limiting its transferability to other tasks. In particular, the audio features are constrained in human speech, leading to poor diversity of audio feature distribution. It further hinders the modelling of cross-model representation.

There also exists some work learning audio-text multimodal representation [10, 11]. However, the domain paradigm for audio-text pre-training is cross-modal contrastive learning, which relies on a dual-encoder to encode data from different modalities and trains the model on giant corpora of audio-text pairs. By aligning cross-modal paired samples in a common hidden space, the representations learned from contrastive learning are semantically representative and thus can be easily generalized into downstream tasks. Nonetheless, a major deficiency of contrastive learning is its vulnerability to batch size variance[10, 12, 13, 14].

2.2 Proof of Concept

To verify the feasibility of the logocentrism unification framework, I implement a proof-of-concept prototype model, which is comprised of a mature backbone network and a pre-trained language model. The prototype model consists of two components: an encoder that encodes spectrogram and a language model decoder that generates language tokens. In my implementation, I apply a ViT [15] as the audio encoder and a pre-trained GPT-Neo [16] as logos reasoner.

Audio Input To better make use of the image understanding ability of ViT [15], different from WavPrompt [8], the PoC model employs log Mel spectrograms as audio features. Specifically, I transform audio clips into Mel spectrograms and divide them into grid patches, which are then embedded by a linear layer and flattened into a sequence of spectrogram tokens. To encode 2D distribution information of patches, following MAE[17], patches are then added with sinusoidal position embeddings.

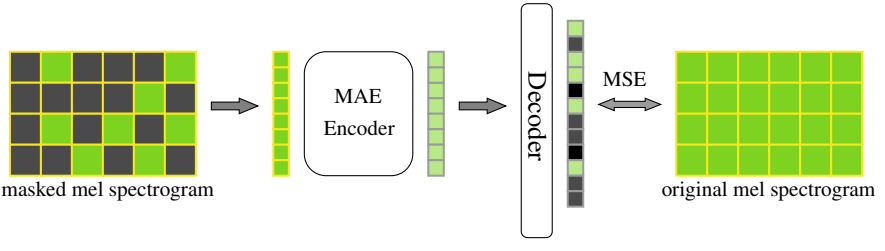


Figure 2: The illustration of the mask training strategy of MAE encoder. Following the regular training pipeline of visual masked auto-encoder, the spectrogram is patched to the spectrogram patch sequence. Then, the 25% patches are fed into the audio MAE to generate the reconstructed spectrogram, and training objective function is defined on the MSE between the reconstructed and original spectrogram. The encoder part of audio MAE is preserved in the next stage.

Audio Encoder The PoC model employee a ViT [15] to extract audio-related embedding from Mel spectrogram, which is a variant of transformer network and has better compatible with the language model. To enhance the ability of audio encoder to learn a general-purpose representation in various audio tasks, the model is pre-trained with masked spectrogram modelling, which is a self-supervised surrogate task and has proved its effectiveness on various downstream tasks [18].

Although MAE masked a large proportion of input patches for reconstruction. In our method, the accuracy of audio-text translation is the first priority. Thus, I mask a small portion of patches to regularize the learning process, which can enhance the model's robustness to disturbance while not impairing its translation precision. As a side effect, the

total computation cost is lower than where all patches are fed. Considering the similarity with language modelling, the mask ratio is set to 15 %, the same as BERT. Following MAE, the encoder processes non-masked patches only, that is 85% of total patches. This strategy reduces computation load, which is quadratic to the sequence length.

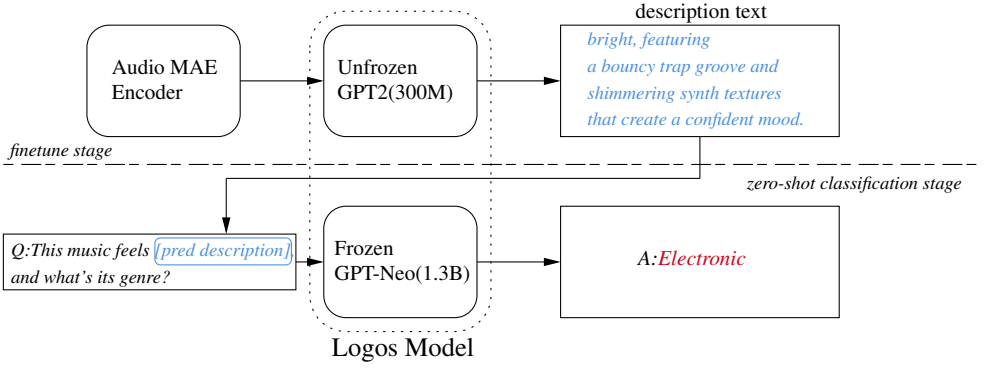


Figure 3: The design of the PoC model. In the finetune stage, the Audio MAE Encoder and unfroze GPT2 are trained in end-to-end manner to predict the music description text with audio inputs. After training, the model generates the description texts for test audio samples. Then, a frozen GPT-Neo is employed to do the zero-shot classification by constructing prompt questions with predicted text and generating the corresponding answer. Finally, the classification result is read from the generated answer.

Logos Model Regarding the design of Logos model, the PoC model utilizes a hybrid strategy to handle the transcription of input audio and the zero-shot knowledge reasoning individually. In detail, a pre-trained GPT-2 [7] model with 300 million parameters is combined with the aforementioned pre-trained audio MAE encoder to form an end-to-end transcription network. This network is trained to predict the description text directly rather than learning a text-aligned embedding, which has the human-readable output and better interpretability for debugging this PoC implementation. Subsequently, the predicted description texts are fed into a pre-trained GPT-Neo [16] with 1.3 billion parameters for zero-shot reasoning. In this stage, I employ the prompt learning technique [19], which constructs a question from input texts with a prompt template, to make the GPT-Neo model do reasoning and return the desired answer. As the Fig.3 shows, if the audio encoder gives the description texts like *"bright, featuring a bouncy trap grove..."*, the prompt learning method constructs a question like *"Q: This music feels bright, featuring a bouncy trap grove..., and what's its genre?"*. In the ideal scenario, the GPT-Neo [16] model will fill the answer part, and this part is extracted for the final genre classification result. It's worth noting that the former GPT-2 model is trainable for fitting the task of description texts prediction, and the GPT-Neo model is frozen from start to finish which has better generalization ability without the potential risk of catastrophic forgetting [20] caused by the finetuning on audio-related data.

2.3 Preliminary Results

For the audio encoder, I reproduce the msm-mae [18], a masked auto-encoder architecture for audio, which learns the representations of audio pieces by reconstructing masked log Mel

Spectrogram. For the audio decoder, I employ a pre-trained GPT2. To get spectrograms, I process samples with a sampling frequency of 16,000 Hz, the window size of 25 ms, hop size of 10 ms, and Mel-spaced frequency bins as 80 in the range 50–8,000 Hz.

The experiments were conducted on music genre classification and sound event detection tasks independently, to investigate the generalization of the PoC model on different task across audio and music fields.

Prompts	language model	0-shot Acc
The music feels <des>. So its genre is	GPT2-Large	22.7
	GPT-NEO-1.3B	32.3

Table 1: Top-1 accuracy of zero-shot genre classification results on FMA Small dataset.

Prompts	Language model	Train Data Scale	Test Set	Acc
The music feels <des>. So its genre is	GPT-NEO-1.3B	22K	GTZAN	44.3
			FMA Small	32.3
		1.5M	GTZAN	64.8
			FMA Small	31.8

Table 2: Top-1 accuracy of zero-shot genre classification results on FMA Small and GTZAN dataset.

For the zero-shot music genre classification task, I first pre-train a PoC model and then employ its music understanding ability for downstream zero-shot tasks. For pre-training, I collect 28,583 music-text pairs from ShutterStock where every piece of music has corresponding description text. According to our observation, the composition of description texts follows a fixed routine: begin with two adjective words, follow music features and elements, and end with moods reflected by the audio. This stable pattern enables our model to learn similar music comments effectively. It also facilitates the construction of divided subdatasets. Regarding data split, I randomly selected 25,723 music-text pairs as the training set and 1,430 for the development and test set respectively. Then I pre-trained our PoC model on this corpus for 100 epochs and choose the checkpoint with the best testing set performance for genre classification.

The zero-shot experiments were conducted on fma small [21] and GTZAN. As described in Sec 2.2, descriptive texts generated by our pre-trained PoC are wrapped by a prompt-formulation function, *'The music feels* ⟨des⟩*. So its genre is'* . Then the whole prompt sentence is feed into a GPT-Neo for inference. Since the label number is limited, I select the label token with the highest logit value as a prediction. The results are shown in Tab. 1. Zero-shot performance surges with the scaling up of the language model, from 22.7 to 32.3. For further analysis, Fig. 4 represents the confusion matrix. I found that language models are biased, but the increase in their scale can alleviate this issue. The prediction of GPT2-large is highly biased towards the genre 'hip-hop'. However, GPT-Neo, which is pre-trained by a more extensive and more comprehensive corpus, treats every genre more equally. As shown on the matrix diagonal, the accuracy of some genres like 'folk' and 'electronic' even exceeds 50%. Another observation is that for labelling words with multi meanings, such as 'experimental' or 'international', GPT models can not understand them well. Besides, after collecting more data, I compared the zero-shot learning performance of models trained on different amounts of data. I found that training on 1.5 million pairs music-text data significantly improves the performance of GPT-3 on GTZAN, but interestingly, the performance

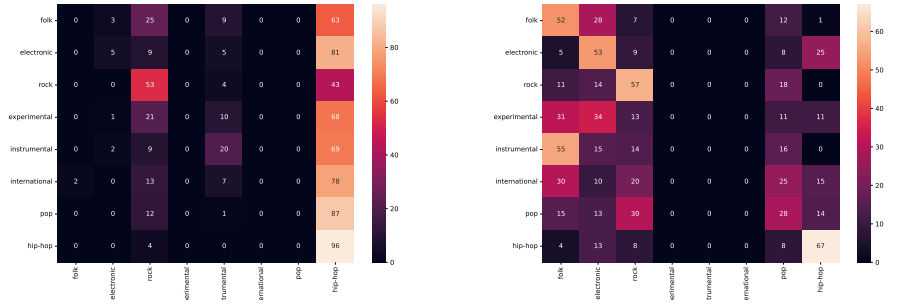


Figure 4: Confusion matrix of zero-shot results on FMA Small. The left matrix is from GPT-large reuslts. The right figure presents results from GPT-Neo-1.3B

on FMA small actually decreased slightly, as Tab. 2 shows. My initial hypothesis is that the music-language pretrained model is highly sensitive to the prompt text, so on different test sets, the model needs to use completely different prompts to achieve optimal performance. This is an interesting question that deserves further exploration.

Prompts	language model	Performance
The audio feels like	GPT2-Base	24.3
	GPT2-Large	25.7

Table 3: mAP on Fully supervised results on AudioSet

I also evaluated our PoC model's ability of sound event detection on AudioSet [22]. Considering the event description words are constrained in limited label space, I conducted fully supervised training and testing. Rather than a two-step inference mechanism, I predict the label words directly after the prompt 'The audio sounds like '. Results are shown in Tab. 3 demonstrate that my model can also handle multi-label tasks.

3 Open Problems and Work Plan

3.1 Open Problems

The implementation and validation of the logocentrism framework yield initial results that prove the feasibility of this design. In particular, the zero-shot learning accuracy on genre classification task is promising and indicates the potential of generalization for more audio and music-related tasks. Nevertheless, the implementation still has a long way to go in developing a truly unified system that performs well on all tasks. Based on observation and rethinking on the current work, the remaining research problems can be categorized as following:

- **Self-supervised representation Learning of Audio** The representation learning problem is critical to a model that aims to solve different tasks in a single model. A unified framework naturally requires the representation of input is task-agnostic. The advantage of self-supervised learning in this problem is self-evident due to the irrelevance between self-supervised pretext tasks and downstream tasks. However, self-supervised

learning methods for speech have often been borrowed from the computer vision domain. For example, masked spectrogram modelling is the audio variant of masked image modelling, and the audio contrastive learning is literally the clone of image contrastive learning with nuances of input and augmentation methods. Nevertheless, the time-frequency representation of audio has its intrinsic characteristics which are different from image. A difference that is often neglected is that the spectrograms do not have translation invariance like images due to the physical meaning of every time-frequency bin. Thus, it's valuable to explore the self-supervised learning method that takes account of the properties of audio and its spectrogram representation.

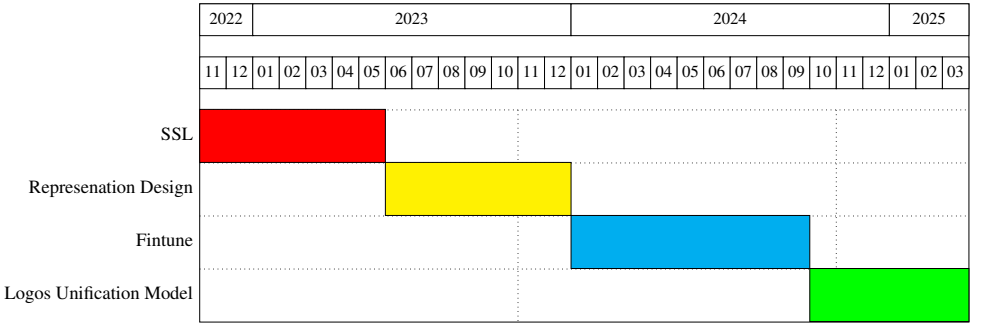
- **Design of Audio Representation** In the PoC model, Audio Representation is dense feature map which is extracted by a ViT-based model. Meanwhile, there is other literature [8] that suggests using the discrete representation that is generated by quantization techniques [23]. Each of these two manners has its own advantages. The dense way is more straightforward and easy to implement. And discrete representation can be modelled directly by auto-regressive model with sampling operation. A hybrid representation form may own the advantages of both methods. The design of audio representation is also an interesting problem in the field of audio-text joint modelling.

- **Dense Prediction Problem** Depending on the structure of transformer and the auto-regressive computation pattern, the large language models prefer to yield discrete outputs rather than dense outputs intrinsically. However, there are many tasks that require models to generate dense prediction, such as speech enhancement, source separation, melody extraction, and audio generation. A auto-regressive transformer-compatible dense prediction mechanism is worth investigating.

- **Finetune and Continual learning of language model** The LLMs are usually pre-trained on enormous corpora which is collected from various sources on web. The audio and music-related content take up a small proportion of the whole data. Thus, normal LLMs can not generalize well on audio tasks. To alleviate this problem, the PoC model pre-trains the GPT-2 model on audio-text paired data. However, the finetune of pre-trained model causes the bothersome catastrophic forgetting problem, and the finetune of LLMs on audio-related data is no exception. The catastrophic forgetting impair the zero-shot generalization ability of LLMs seriously. In the design of PoC model, this problem is bypassed by the introduction of another frozen GPT-Neo Model to do zero-shot learning. It's a temporal fix for demo use rather than an elegant solution. Continual learning has widespread application in resolving this problem. But its effectiveness on the finetune on audio-text data has not been investigated.

3.2 Work Plan

By summarizing the research topics, the work plan can be arranged as following Gantt chart shows.



The main tasks in the work plan include the development of self-supervised learning model for audio (SSL), audio semantic representation design, finetune on audio-text data and logocentrism unification framework.

References

[1] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE, 2022.

[2] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Zero-shot audio source separation through query-based learning from weakly-labeled data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4441–4449, 2022.

[3] Wei-Tsung Lu, Ju-Chiang Wang, Minz Won, Keunwoo Choi, and Xuchen Song. Spectnt: A time-frequency transformer for music audio. *arXiv preprint arXiv:2110.09127*, 2021.

[4] Xingjian Du, Bilei Zhu, Qiuqiang Kong, and Zejun Ma. Singing melody extraction from polyphonic music based on spectral correlation modeling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245. IEEE, 2021.

[5] Xingjian Du, Zhesong Yu, Bilei Zhu, Xiaoou Chen, and Zejun Ma. Bytecover: Cover song identification via multi-loss training. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 551–555. IEEE, 2021.

[6] Zhang Longxi. The" tao" and the" logos": Notes on derrida's critique of logocentrism. *Critical Inquiry*, 11(3):385–398, 1985.

[7] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[8] Heting Gao, Junrui Ni, Kaizhi Qian, Yang Zhang, Shiyu Chang, and Mark Hasegawa-Johnson. Wavprompt: Towards few-shot spoken language understanding with frozen language models. *arXiv preprint arXiv:2203.15863*, 2022.

[9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. Technical report, Technical report, OpenAI, 2022. URL https://cdn. openai. com/papers/whisper. pdf, 2022.

[10] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: Learning audio concepts from natural language supervision. *arXiv preprint arXiv:2206.04769*, 2022.

[11] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[13] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[14] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[16] Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. If you use this software, please cite it using these metadata.

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[18] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. *arXiv preprint arXiv:2204.12260*, 2022.

[19] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021.

[20] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

[21] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

[22] 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 *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[23] Robert Gray. Vector quantization. *IEEE Assp Magazine*, 1(2):4–29, 1984.