MAP-MUSIC2VEC: A SIMPLE AND EFFECTIVE BASELINE FOR SELF-SUPERVISED MUSIC AUDIO REPRESENTATION LEARNING

Yizhi Li[1*] **Ruibin Yuan**[2,4*] **Ge Zhang**[2,5*] **Yinghao Ma**[3*]
Chenghua Lin[1†] **Xingran Chen**[5] **Anton Ragni**[1] **Hanzhi Yin**[4] **Zhijie Hu**[6]
Haoyu He[7] **Emmanouil Benetos**[3] **Norbert Gyenge**[1] **Ruibo Liu**[8] **Jie Fu**[2†]

[1]Department of Computer Science, University of Sheffield, UK
[2]Beijing Academy of Artificial Intelligence, China
[3]Centre for Digital Music, Queen Mary University of London, UK
[4]School of Music, Carnegie Mellon University, PA, USA
[5]University of Michigan Ann Arbor, USA
[6]HSBC Business School, Peking University, China
[7]University of Tübingen & MPI-IS, Germany
[8]Dartmouth College, NH, USA
{yizhi.li, c.lin}@sheffield.ac.uk, yinghao.ma@qmul.ac.uk, fujie@baai.ac.cn

ABSTRACT

The deep learning community has witnessed an exponentially growing interest in self-supervised learning (SSL). However, it still remains unexplored how to build a framework for learning useful representations of raw music waveforms in a self-supervised manner. In this work, we design Music2Vec, a framework exploring different SSL algorithmic components and tricks for music audio recordings. Our model achieves comparable results to the state-of-the-art (SOTA) music SSL model Jukebox, despite being much smaller with less than 2% of parameters of the latter. The model will be released on Huggingface [1] .

1. INTRODUCTION

SSL has been proven effective for extracting features from raw music waveforms [1, 2]. Unfortunately, existing models (e.g., Jukebox [1]) are prohibitively expensive to fine-tune and extend to different applications [2] . As an effort to obtain the computationally affordable baseline, we design and train Music2Vec. It mainly follows the design principles proposed in data2vec [3]. Our key contributions are as follows: (1) developing music2vec, an open source self-supervised system for raw music files with single-GPU trainable size (about 90M parameters); (2) demonstrating that the model achieves comparable results to Jukebox on multiple music information retrieval tasks.

2. METHOD

We follow the standard pretraining protocols of data2vec [3] with the fairseq framework [4], and further release our computationally affordable models.

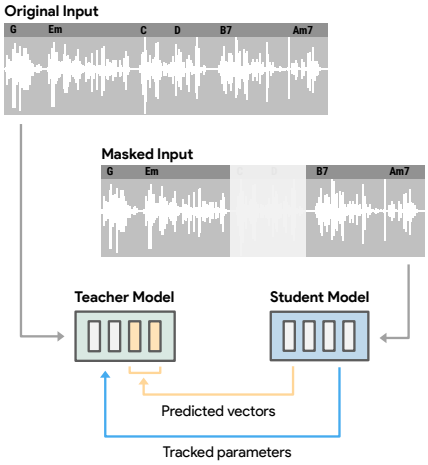

Figure 1. Music2Vec Framework. During pre-training, the student model aims to reconstruct the masked music audio by taking the contextualised representations provided by teacher model as prediction target.

Data2vec claims a unified SSL framework for either speech, NLP, or computer vision following the design of bootstrapping learning [5], which is illustrated in Fig. 1. The teacher model and student model share the same architecture, and the parameters of the teacher model are updated according to the exponential moving average of the student [3]. The student model takes the partially masked input and is asked to predict the average of top-K layer outputs of Transformer in the teacher model. In contrast, the teacher model takes the unmasked input and provides contextual prediction targets in the pre-training.

We directly apply the Data2Vec base model, which encodes audio recordings using a multi-layer 1-D CNN feature extractor mapping 16 kHz waveform to 50 Hz representations [13], and further input the encoded tokens into a 12-layer Transformer Blocks with $H = 768$ hidden dimension (with $4\times H$ feed-forward inner-dimension). Since starting with pre-trained speech models can barely benefit music representation learning [14], we instead train the randomly initialised base model from scratch to verify its effectiveness on modelling music audio recordings.

We collect around 130k hours of music audio files from

* The authors contributed equally to this work.
† Corresponding authors.
[1] Please refer to our huggingface checkpoint.
[2] Jukebox needs over 10GB to store activations.

Late-Breaking / Demo Session Extended Abstract, ISMIR 2022 Conference

	Approach		Tags (MTT [6]) AUC	Tags (MTT [6]) AP	Genre (GTZAN [7]) Accuracy	Key(GS [8]) Score	Emotion (EMO [9]) $R2_{Arousal}$	Emotion (EMO [9]) $R2_{Valence}$	Average
Baselines	CHOI [10]		89.7	36.4	75.9	13.1	67.3	43.4	51.9
	MUSICNN [11]		90.6	38.3	79.0	12.8	70.3	46.6	53.7
	CLMR [12]		89.4	36.1	68.6	14.9	67.8	45.8	50.8
	Jukebox [1]		91.5	41.4	79.7	66.7	72.1	61.7	69.9
Music2Vec	Starting Setting		88.2	34.1	61.7	32.1$^{\diamond}$	66.2	45.8	54.7
	Length Crop	5s	89.5	35.9	76.6	50.1$^{\diamond}$	69.4	57.4	63.2
		10s	89.0	36.0	70.3	27.4	62.7	46.1	55.3
		15s	88.3	34.1	65.9	38.1$^{\diamond}$	60.1	43.6	55.0
	Mask Span	5	87.0$^{\diamond}$	32.2$^{\diamond}$	59.3	29.5	50.3$^{\diamond}$	24.7$^{\diamond}$	47.2
		15	87.8	33.3	65.2	41.9$^{\diamond}$	55.0	36.9	53.4
	Mask Prob	50%	87.7	33.2	62.8	43.6$^{\diamond}$	54.8	37.6	53.2
		70%	87.2	32.4	60.7	35.3$^{\diamond}$	55.3$^{\diamond}$	36.0$^{\diamond}$	51.2
		80%	87.5	32.7	60.0$^{\diamond}$	34.6$^{\diamond}$	50.7	40.4	51.0
	Target	Top-12	88.8	34.5	65.2	50.8$^{\diamond}$	67.4	43.8	58.4
	Step	800K	87.6$^{\diamond}$	33.2$^{\diamond}$	60.3	44.9$^{\diamond}$	54.8$^{\diamond}$	40.8$^{\diamond}$	53.6

Table 1. Overall Results of Self-Supervised Models. We report the results of Music2Vec trained with controlled variables derived from the speech data2vec setting [3]. Underline and square box indicate the best overall performance and the best setting of Music2Vec, respectively. $\diamond$ indicates the results are produced by the convolutional feature extractor representations. We use dots with different colors to present different hyperparameters: length crop, mask span, mask prob, target, and step. Results of baselines are taken from JukeMIR [2] and datasets for different tasks are given in brackets.

the Internet and use a 1k hours subset that contain 30s long wave files to train our model. All Music2Vec models are trained for 400k steps with 8 $\times$ NVIDIA A100-40GB GPUs. Training with eight GPUs takes around 6 days, i.e., about 48 days with only one A100 GPU.

3. EXPERIMENTS

3.1 Dataset and Evaluation

We follow the probing evaluation setting of JukeMIR [2] to verify the music modelling performance of our models. Specifically, we report results on a comprehensive set of music information retrieval tasks, including multi-label **tagging**, multi-class **genre classification**, multi-class **key detection**, as well as a regression task **emotion recognition**. Following the evaluation setting [2] for the SOTA pre-trained model, AUC (the area under the receiver operating characteristic curve) is regarded as the main metric of tagging to select checkpoints, and the marco average of arousal and valence R2 decides for emotion recognition.

3.2 Pre-train Settings

Adapting from the data2vec model on auditory signals [3], we conduct parameter searching and correlation analysis for Music2Vec pretraining, including the recording length, the mask strategy, and the learning target layers.

First, we use **audio length cropping** to shorten music excerpts, since longer sequences are more difficult for modelling. Note that the combined audio file length in a batch is not altered and the hardware environment remains the same, which makes a single training batch contains larger number of music samples when cropping the clips.

Second, we revise the mask strategy by changing **mask span length** and **mask token probability**. Mask token probability is the probability for each token to be chosen as the start of the span to be masked, and the length of which can also be adapted for different data modalities [3].

Third, we modify the **prediction target** provided by the teacher model. Our preliminary experiments illustrate that early layer representations generally perform well on key detection. Therefore, we change the prediction target in Music2Vec from the average of the top-8 layer representations to all the 12 layers, so that the student model might benefit from the potentially preserved key information.

[3] We use audio files with 30s length, mask span length 10, mask probability 65%, target top-8, and training step 400K as the starting setting. The results is shown as the starting setting in the table.

4. RESULTS AND CONCLUSION

From Tab. 1 we observe that the Music2Vec with the best setting (i.e., crop5s) achieves comparable results to Jukebox on music information retrieval tasks with less than 1/50 parameters of the latter. The audio file length is negatively correlated to the Music2Vec performance, which implies that modelling long sequence is still challenging.

Noticeably, the CNN representations sometimes outperform the Transformer layers, especially for key detection. When including extra early Transformer layers to the prediction target, Music2Vec achieves performance gains in most tasks. This implies that with little or no contextualisation our model still manages to perform fairly well with local features (similar to bag-of-words). However, this also suggests that our model relies too much on local information and leaves a large room for improvement when taking long-range contextual information into consideration. Last but not least, we find that increasing the training steps, changing the mask span, or changing the mask probability does not give performance gain in most tasks.

In conclusion, we propose a training framework for music audio recording pre-training on large-scale data, which gives comparable performance to SOTA models. Our framework also has great potential for efficient fine-tuning and model distillation, which we leave for future work.

5. ACKNOWLEDGEMENTS

This paper is a tribute to our talented friend Anqiao Yang, for his friendship and valuable advice to this work. Yizhi Li is fully funded by an industrial PhD studentship (Grant number: 171362) from the University of Sheffield, UK. Yinghao Ma is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1]. This work is supported by the National Key R&D Program of China (2020AAA0105200). We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing. We would also like to express great appreciation for the suggestions from faculties Dr Chris Donahue, and Dr Roger Dannenberg, as well as the facility support from Mr. Yulong Zhang in the preliminary stage.

6. REFERENCES

[1] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[2] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," *arXiv preprint arXiv:2107.05677*, 2021.

[3] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[4] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[5] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[6] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging." in *ISMIR*, 2009, pp. 387–392.

[7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[8] P. Knees, Á. Faraldo Pérez, H. Boyer, R. Vogl, S. Böck, F. Hörschläger, M. Le Goff *et al.*, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR); 2015 Oct 26-30; Málaga, Spain.[Málaga]: International Society for Music Information Retrieval, 2015. p. 364-70.* International Society for Music Information Retrieval (ISMIR), 2015.

[9] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM international workshop on Crowdsourcing for multimedia*, 2013, pp. 1–6.

[10] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Transfer learning for music classification and regression tasks," *arXiv preprint arXiv:1703.09179*, 2017.

[11] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," *arXiv preprint arXiv:1909.06654*, 2019.

[12] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," *arXiv preprint arXiv:2103.09410*, 2021.

[13] 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.

[14] A. Ragano, E. Benetos, and A. Hines, "Learning music representations with wav2vec 2.0," *arXiv preprint arXiv:2210.15310*, 2022.