ON THE EFFECTIVENESS OF SPEECH SELF-SUPERVISED LEARNING FOR MUSIC

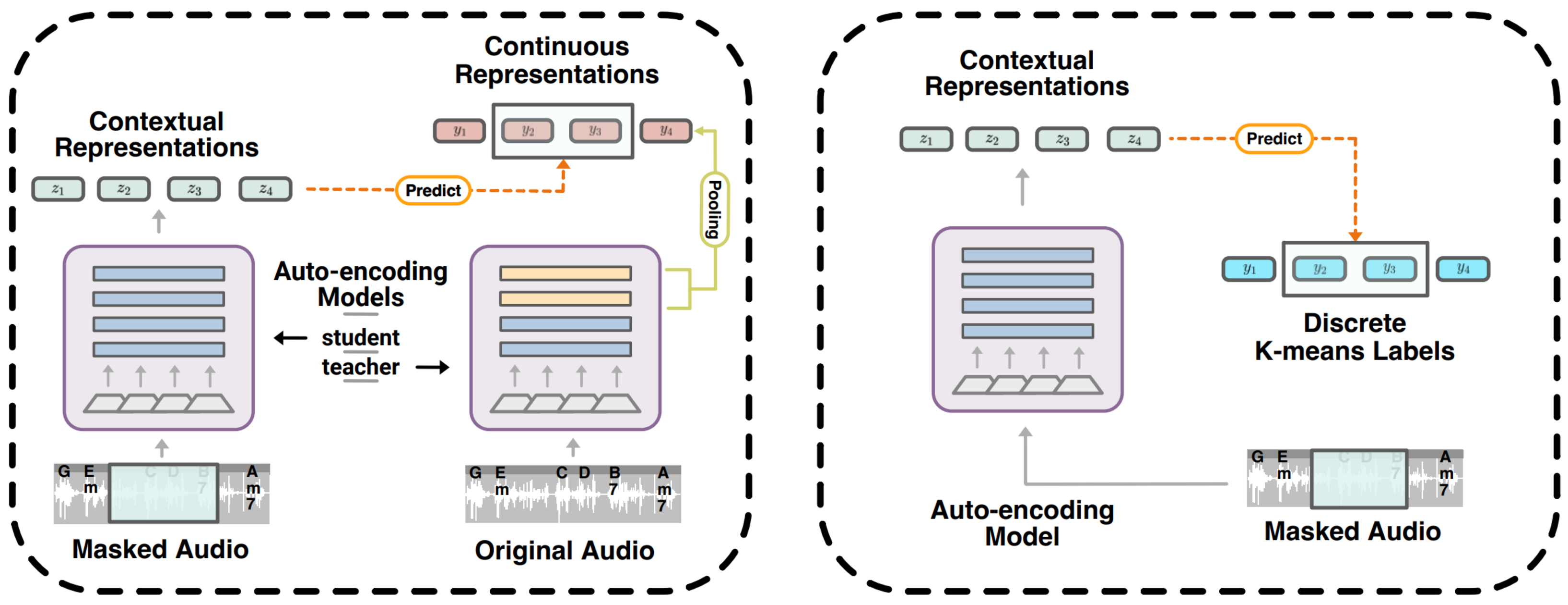
Yinghao Ma[1*] Ruibin Yuan[2,3*] Yizhi Li [4*] Ge Zhang[2,5*] Xingran Chen[6] Hanzhi Yin[3] Chenghua Lin[4†] Emmanouil Benetos[1†] Anton Ragni[4] Norbert Gyenge[4] Ruibo Liu[7] Gus Xia[8] Roger Dannenberg[3] Yike Guo[9] Jie Fu[2†]

1 Queen Mary University of London 2 Beijing Academy of Artificial Intelligence 3 Carnegie Mellon University 4 University of Sheffield 5 University of Waterloo 6 University of Michigan Ann Arbor 7 Dartmouth College 8 New York University Shanghai 9 Hong Kong University of Science and Technology {yinghao.ma, emmanouil.benetos}@qmul.ac.uk, ruibiny@andrew.cmu.edu {yizhi.li, c.lin}@sheffield.ac.uk, gezhang@umich.edu, fujie@baai.ac.cn

1. Introduction

Self-supervised learning (SSL) has shown promising results in speech, but its efficacy in music information retrieval (MIR) still remains largely unexplored.

- Applying open-source speech SSL models (data2vec1.0[1] and HuBERT[2]) to music recordings, referring as **Music2vec** and **MusicHuBERT**, respectively.
- We train 12 SSL models with 95M parameters under 13 different MIR tasks.
- Identifying weaknesses for further research.



(a) Music2Vec (b) MusicHuBERT

Figure 1: Pre-training Paradigms of Selected Models. Both of the models are fed with masked audio inputs and predict given targets without supervised information.

2. Methodology

- Masked language model.
- Music2Vec: use a **teacher model** in the same architecture to provide deep features for prediction targets in the reconstruction of masked audio.
- MusicHubert: **k-means** clustering results for **MFCCs** features of music audio as reconstruction targets.
- Architecture: a multi-layer **1-D CNN feature extractor**, and further input these tokens to a 12-layer Transformer with dimension 768.
- We trained these models on 1k hours of 5s music audio, with 8 × NVIDIA A100-40GB GPUs around2- 3 days for 250k steps.

3. Pre-training Experiments

- Training dataset
- K-means for MusicHubert
- Prediction Target of Music2Vec

Discussion of Table 1 ↓

- MusicHuBERT surpasses Music2vec in various tasks
- Pre-training with HuBERT is strongly linked to MFCC features, limiting multi-pitch information
- Music2Vec is better at learning pitch information, but worse at beat tracking.

Table 1: Experimental performance of the SSL baseline systems on all downstream tasks

Downstream dataset	MTT		GS key	GTZAN Genre	EMO		Nsynth Instr	Nsynth pitch	VocalSet tech	VocalSet singer	GTzAN Rhythm	MTG Instrument		MTG MoodTheme		MTG Genre		MTG Top50	
Metrics	ROC	AP	Refined Acc	Acc	Emo_V	Emo_A	Acc	Acc	Acc	Acc	F1 (beat)	ROC	AP	ROC	AP	ROC	AP	ROC	AP
HuBERT base	89.8	36.4	15.0	64.8	31.0	57.5	68.2	79.4	61.0	58.8	83.5	73.2	17.0	74.0	11.6	85.0	16.3	81.8	26.5
MusicHuBERT base	90.2	37.7	14.7	70.0	42.1	66.5	69.3	77.4	65.9	75.3	**88.6**	75.5	17.8	76.0	13.9	86.5	18.0	82.4	28.1
data2vec audio base	88.4	33.6	15.5	60.7	23.0	49.6	69.3	77.7	64.9	74.6	36.4	73.1	16.9	73.3	11.0	83.5	14.5	80.6	24.8
Music2vec vanilla	89.1	35.1	19.0	59.7	38.5	61.9	69.4	88.9	**68.3**	69.5	33.5	73.1	16.3	74.3	12.2	85.2	16.5	81.4	26.2
SOTA	**92.0** [40]	**41.4** [6]	**74.3** [28]	**82.1** [41]	**61.7**	**72.1** [6]	**78.2** [20]	**89.2** [23]	65.6 [36]	**80.3** [42]	80.6 [43]	**78.8**	**20.2** [44]	**78.6**	**16.1** [23]	**87.7**	**20.3** [44]	**84.3**	**32.1** [23]

Table 2: Ablation study on MusicHuBERT hyperparameters (k is the number of MFCC clusters)

Downstream dataset	MTT		GS key	GTZAN Genre	EMO		Nsynth Instr	Nsynth pitch	VocalSet tech	VocalSet singer	GTZAN Rhythm	Average Score
Metrics	ROC	AP	Refined Acc	Acc	Emo_V	Emo_A	Acc	Acc	Acc	Acc	F1 (beat)	score
HuBERT	89.8	36.4	15.0	64.8	31.0	57.5	68.2	79.4	61.0	58.8	83.5	59.8
k=2000 MFCC dim=39	90.2	37.7	14.7	**70.0**	42.1	66.5	69.3	77.4	65.9	75.3	88.6	64.4
k=2000 iter2	**90.4**	37.5	13.8	68.3	**43.3**	67.4	70.0	**80.3**	63.6	70.4	**88.8**	63.8
k=500 MFCC dim=39	89.6	36.1	15.7	64.5	41.0	67.7	66.7	76.8	60.5	72.3	87.5	62.4
k=500 MFCC dim=60	90.3	**38.0**	**17.6**	69.7	40.8	**67.5**	**70.3**	79.0	**66.2**	**75.5**	88.6	**65.0**

Discussion of Table 2 ↑

- MusicHuBERT with k=2000 outperforms k=500 for most tasks
- K-means clustering of deep features performs better than vanilla MusicHuBERT for most tasks, except pure vocal datasets.
- Increasing the dimension of MFCC doesn't significantly impact most tasks.

Discussion of Table 3 →

- Modifying the prediction target for Music2Vec from the average of the top 8 layers to all 12 layers enhances performance across various tasks, notably improving key detection.
- The use of audio length cropping for shorter music excerpts is introduced to ease modelling difficulties with longer sequences, revealing that key detection results may be affected by local versus global key differences in shorter segments.

Table 3: Ablation study on Music2Vec hyperparameters (span is mask span, prob is mask probability, step is training steps, target=12 uses all 12 transformer layers, and crop5s uses 5s music excerpts)

Downstream dataset	MTT		GS key	GTZAN Genre	EMO		Nsynth Instr	Nsynth pitch	VocalSet tech	VocalSet singer	GTZAN Rhythm	Average Score
Metrics	ROC	AP	Refined Acc	Acc	Emo_V	Emo_A	Acc	Acc	Acc	Acc	F1 (beat)	score
data2vec	88.4	33.6	15.5	60.7	23.0	49.6	69.3	77.7	64.9	**74.6**	**36.4**	55.2
vanilla	89.1	35.1	19.0	59.7	38.5	61.9	69.4	88.9	68.3	69.5	33.5	57.8
span=5	87.3	32.0	15.7	47.6	22.7	41.2	64.2	84.8	56.7	53.8	33.2	49.7
span=15	88.7	34.3	16.4	56.6	39.0	58.8	67.1	88.1	63.1	61.9	33.1	55.2
prob=50	88.5	34.0	23.7	59.3	40.6	55.0	66.8	87.7	64.9	61.7	33.9	56.3
prob=80	88.2	33.9	18.4	50.3	36.7	55.7	67.9	88.9	64.2	65.2	33.7	55.1
step=800k	87.7	32.7	20.3	54.5	34.9	47.3	66.9	87.5	65.6	65.1	33.4	55.0
target=12	89.7	35.2	**26.5**	64.5	41.7	64.2	**71.1**	**89.2**	71.0	73.2	34.1	60.6
crop5s	**90.0**	**36.6**	18.5	**76.6**	**53.4**	**71.6**	68.3	88.9	**71.3**	72.4	33.9	**61.8**

Music2Vec & MERT Model Released Here

4. Results

- Pre-training with music recordings rather than speech can generally improve performance on a wide range of MIR tasks, even when the models and training are designed for speech.
- some limitations and suggestions for the following pre-training:
  - > emphasis key or harmonic by replacing MFCC features
  - > larger number "k" for k-means compared to speech phones.
  - > different "k" for pitch and timbre.
- Shortening the audio length can Increase batch diversity, providing better performance.
- An improved pre-trained model **MERT** https://arxiv.org/abs/2306.00107

References

[1] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for selfsupervised learning in speech, vision and language," arXiv preprint arXiv:2202.03555, 2022.
[2]] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia,
R. Salakhutdinov, and A. Mohamed, "Hubert: Selfsupervised speech representation learning by masked
prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451–3460, 2021.

Acknowledgements

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]. Yizhi Li is fully funded by an industrial PhD studentship (Grant number: 171362) from the University of Sheffield, UK. 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.