Persian Traditional Music Information Retrieval (Fifth Week Report)

Pouya Mohseni

Supervisor: Dr. Bagher BabaAli

Abstract

Self-supervised learning has shown to be a powerful tool for extracting meaningful information in various domains including speech and music. Learned representations using these models have proven to lead to pleasant results in different downstream tasks in information retrieval. Yet, the use of self-supervised methods is underexploited in the domain of Persian Traditional Music. In this research, we aim to explore the use of these methods in this music domain and address downstream tasks defined exclusively in this paradigm, ultimately contributing to a deeper understanding and preservation of this culturally significant music heritage.

1 Introduction

Deep learning methods have brought promising results in various end-to-end tasks, including supervised tasks such as natural language processing (NLP) and computer vision. However, the scarcity of labeled data in various domains including machine listening is one of the challenges facing the development of richer models. To tackle this problem, self-supervised learning (SSL) methods have been employed for training models. These methods rely on leveraging enormous unlabeled data to pre-train a model that can capture data representation and ultimately fine-tune on the limited tagged data.

There are different approaches to SSL for music representation learning. State-of-the-art (SOTA) feature extractors trained on speech domains such as

PANNs [Kong, 2020], PASE [Wu, 2021], HUBERT [Hsu, 2021], or Wev2Vec 2.0 [Baevski, 2020] could be retrained, pre-trained with the learned weights, or fine-tuned on the music domain. Adapting Speech SSL frameworks based on mask prediction to the music domain by pretraining them on the music domain results in promising models in MIR downstream tasks. MAP-MERT v0 [Li, 2022] and MERT [Li, 2023] based on HuBERT, and Mus2Vec [Ma, 2023] based on Wev2vec 1.0 are examples of these approaches. However, only a few of these models are open-sourced, hindering the exploration of their potential in the Persian Traditional Music domain. Adapting Speach SSL models based on instance discrimination is another approach. Models such as CLMR [Spijkervet, 2021] and PEMR [Yao, 2022] are based on this approach. Nevertheless, these models do not achieve competitive results compared to other models.

Persian Traditional Music MIR has many advantages, serving as a tool for preservation, teaching, and recommendation purposes. MIR technologies help preserve the rich cultural heritage of Persian Traditional Music by providing methods for digitizing and archiving records and written notes, which ensures the accessibility of future generations and the spread of this music. Moreover, MIR facilitates teaching and learning music by adapting old scores to new technologies for transcription, analysis, and visualization. Lastly, MIR aids in recommending this kind of music to enthusiasts by considering music characteristics as well as individual preferences. This, in turn, will lead to further commercializing Persian Traditional Music and enticing more artists and individuals to explore this kind of music. By harnessing the power of MIR, Persian traditional music can continue to thrive, evolve, and inspire audiences worldwide.

2 Background and Related Work

2.1 Persian Traditional Music Datasets

Various datasets have been created to pave the way for MIR research in the domain of Persian Music. To the best of our knowledge, the first attempt to construct a dataset for Dastgah classification in Traditional Persian Music was undertaken by in [Heydarian, 2005] which captures tracks from different Dastgah types performed on a Santur, with a total duration of approximately 100 minutes [Heydarian, 2016]. Proposed in [Abdoli, 2011], the relatively small dataset for Persian Traditional Music comprises 210 tracks representing various Dastgah types from different artists. Another balanced-class dataset, focusing on instrument recognition in Traditional

Persian Music, was created by [Mousavi, 2019], including about 700 samples of 5-10 second tracks from seven traditional instruments. The dataset utilized in [Geravanchizadeh, 2022] contains 250 solo pieces by M. Alizadeh, equally distributed among five Dastgah types. Another dataset proposed in [LAYEGH, 2013]addresses the Radif-e Mirza Abdollah classification problem, which is specific to Persian Traditional MIR. This unbalanced distributed class dataset encompasses 1250 music samples in 12 modal systems of Dastgah and Avaz. Taking a broader perspective, PMG-Data [Farajzadeh, 2023] consists of 500 tracks from various genres, including Traditional, Pop, Rap, and Monody. To tackle the instrument-independent task of Dastgah classification, the Maryam Iranian classical music (MICM) dataset is proposed [RezezadehAzar, 2018]. MICM comprises 1137 music samples in all seven Dastgah types, although with an unbalanced distribution.

In an effort to advance research in the domain of Persian Traditional Music, the Nava dataset was introduced in [Baba Ali, 2019]. Comparably, Nava is a large and well-balanced dataset, featuring 55 hours of track recordings in 7 Dastgah types and 5 instruments. Moreover, each record is tagged with the instrument name, Dastgah type, and the artist for related downstream tasks. As for the advantages offered, Nava is employed in this research for model training and evaluation in downstream MIR tasks. Nava has been extended in recent studies by adding unlabeled data. The unlabeled section of NAVA consists of 16895 pieces by 181 artists – approximately 960 hours – of Persian Traditional Music [Shirmardi, 2022, Hemati, 2022]. Leveraged by the extended version of NAVA, having self-supervised models pre-trained on Persian Traditional Music is facilitated.

Persian Traditional Music MIR, although advantageous, is relatively underdeveloped. With no universally accepted dataset within the community comparing results of the proposed models in this terrain is impossible. Furthermore, the scarcity of studies in the literature has prohibited SOAT models in MIR from being utilized in this context.

2.2 Persian Traditional Music MIR Approaches

There have been limited studies on Persian Traditional Music MIR, only a few of which are based on SSL approaches. Most of these studies are concerned with Dastgah classification; however, there have been studies on other downstream tasks defined either only on Persian Traditional Music or music in general. Nevertheless, without a benchmark dataset in the community, discussing achieved scores on downstream tasks is trivial as they are not comparable.

Many studies in this terrain utilize traditional machine learning models,

trained on labeled datasets, to tackle MIR downstream tasks. The method introduced in [Abdoli, 2011] classifies the Dastgah of each recording by considering the similarities between the Interval Type 2 Fuzzy Sets (IT2FSs) of records with Dastgah prototypes. In another study [Mousavi, 2019], a neural network-based model is proposed for Persian Classical Music instrument recognition fed on a combination of audio signal spatial and frequency domain features. An SVM approach trained on Lagrange coefficients of pitch logarithm (LCPL) and Fuzzy similarity sets type 2 (FSST2) is proposed in [Geravanchizadeh, 2022] to address Dastgah Classification. Similarly, the method introduced in [LAYEGH, 2013] is based on spatial and frequency domain features classified by a Radif of Mirza Abdollah.

On the other hand, some recent papers leverage deep learning models, trained either with supervised or self-supervised approaches. For instance, AzarNet [RezezadehAzar, 2018], is a deep network that is trained and tested on Short-Time Fourier Transform (STFT) features extracted from samples for Dastgah classification. The idea of training SSL models on Persian Traditional Music has been explored in [Shirmardi, 2022, Hemati, 2022] by discussing and proposing constructive learning approaches. Despite these efforts, it is worth noting that the application of SSL models in the context of Persian Traditional MIR has been relatively limited. This research represents the first comprehensive attempt to explore different approaches to employ SSL models in this domain, aiming to study the potential of these models and contribute to the advancement of the field.

2.3 Self-Supervised Learning in Speech Processing

As speech and music processing deal with the same format and encounter similar challenges, such as the "cocktail party" problem [Brown, 2022], utilizing and adopting speech SOTA models for music processing can lead to promising results in the music processing domain. Many recent advancements in speech processing have been driven by transformer-based architectures [Mohamed, 2022]. And in turn, adopting these models in the music domain has led to pleasant results in MIR downstream tasks and music generation [Ma, 2023, Li, 2023].

Recent advancements in large-scale language models based on mask prediction strategies have found applications in speech processing and subsequently, music processing. The adaptation of the BERT model in the NLP domain to HuBERT in speech processing, and MusicHuBERT or MERT in music processing testifies to this assertion. MusicHuBERT produces pseudo labels for masked audio by resorting to the K-means algorithm that is trained on the Mel-frequency cepstral coefficients (MFCC) features. In other words,

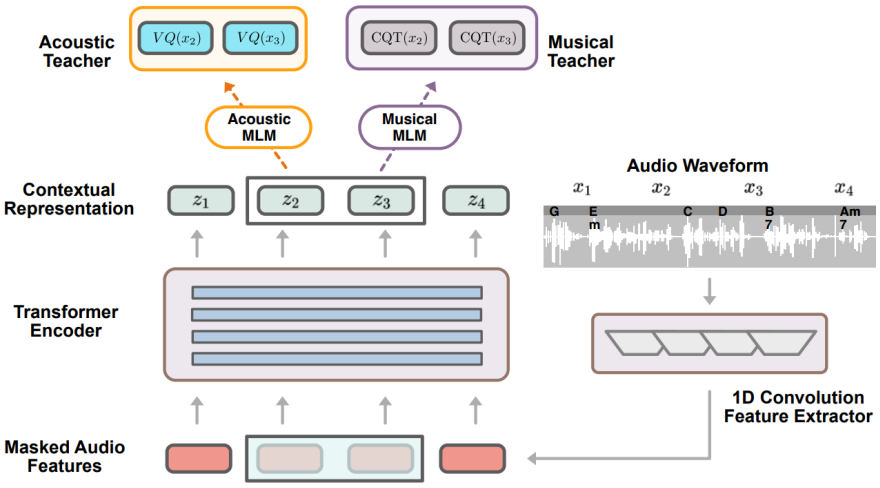

Figure 1: Adapted from [Li, 2023]. Illustration of the MERT Pre-training Framework.

discrete targets are provided by the K-means embeddings of the masked section.

2.4 MERT

In a similar fashion, MERT is based on HuBERT; however, it introduces multi-task learning by providing two types of pseudo labels. An acoustic teacher models acoustic and timbre information and is based on EnCodec [Défossez, 2022]. EnCodec is an 8-layer residual VQ-VAE that converts 24kHz waveforms into 8 embeddings at 75Hz, enabling authentic timbre reconstruction. To emphasize pitch-level information a musical teacher is also introduced. By incorporating Constant-Q transform (CQT) [Brown, 1991] in the reconstruction loss, the model is expected to learn pitch and harmonic inductive bias. Overall, the loss function is proposed as a weighted linear combination of the acoustic-level loss function $\mathcal{L}_H$ and the musical-level loss function $\mathcal{L}_{CQV}$:

$$\mathcal{L}_{totall} = \mathcal{L}_H + \alpha \mathcal{L}_{CQV}$$

Furthermore, to add representation robustness to the model, an in-batch noise mixup is added to MERT. By augmenting audio clips by randomly adding shorter excerpts from the same batch, the model is encouraged to focus on valuable musical sources and ignore noise for improved learning. The MERT architecture is illustrated in Figure 1.

The manifold advantages of MERT, including its multi-task learning approach, acoustic and timbre modeling, and robustness through in-batch noise mixup, collectively make it a compelling choice for our base model for Persian Traditional MIR.

3 Methodology

We adopt three approaches to construct a model addressing Persian Traditional MIR downstream tasks. In the first approach, we employ pre-trained weights from the MERT model on Western music, fine-tuning the model based on the labeled segment of the extended NAVA dataset. Subsequently, we evaluate the model's performance on the predefined downstream tasks of Persian Traditional MIR. In the second approach, we leverage the unlabeled portion of the extended NAVA dataset to pretrain MERT while retaining its learned weights. Then, we assess the model on the downstream tasks by fine-tuning it on the labeled dataset. Lastly, we pre-train MERT from scratch by initializing its weights randomly, pretraining it on the unlabeled segment of the data, fine-tuning it on the labeled segment of the data, and evaluating its performance on the downstream tasks. These three introduced approaches enable us to explore the impact of various training strategies on the performance of MERT in the context of Persian Traditional MIR. The evaluation across predefined downstream tasks will provide insights into the adaptability and effectiveness of each approach in addressing a cross-domain task with MERT.

References

[Kong, 2020] Kong, Qiuqiang, et al. "Panns: Large-scale pretrained audio neural networks for audio pattern recognition." IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020): 2880-2894.

[Wu, 2021] Wu, Ho-Hsiang, et al. "Multi-task self-supervised pre-training for music classification." ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

[Hsu, 2021] Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 3451-3460.

[Baevski, 2020] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in neural information processing systems 33 (2020): 12449-12460.

[Li, 2023] Li, Yizhi, et al. "MERT: Acoustic Music Understanding Model with Large-Scale Self-supervised Training." arXiv preprint arXiv:2306.00107 (2023).

[Ma, 2023] Ma, Yinghao, et al. "On the effectiveness of speech self-supervised learning for music." arXiv preprint arXiv:2307.05161 (2023).

[Li, 2022] Li, Yizhi, et al. "Large-Scale Pretrained Model for Self-Supervised Music Audio Representation Learning." (2022).

[Spijkervet, 2021] Spijkervet, Janne, and John Ashley Burgoyne. "Contrastive learning of musical representations." arXiv preprint arXiv:2103.09410 (2021).

[Yao, 2022] Yao, Dong, et al. "Contrastive learning with positive-negative frame mask for music representation." Proceedings of the ACM Web Conference 2022. 2022.

[Heydarian, 2005] Heydarian, Peyman, and Joshua D. Reiss. "A database for persian music." Proc. of the Digital Music Research Network Summer Conference (DMRN 2005). 2005.

[Heydarian, 2016] Heydarian, Peyman. Automatic recognition of Persian musical modes in audio musical signals. Diss. London Metropolitan University, 2016.

[Abdoli, 2011] Abdoli, Sajjad. "Iranian Traditional Music Dastgah Classification." ISMIR. 2011.

[Mousavi, 2019] Mousavi, Seyed Muhammad Hossein, VB Surya Prasath, and Seyed Muhammad Hassan Mousavi. "Persian classical music instrument recognition (PCMIR) using a novel Persian music database." 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE). IEEE, 2019.

[Geravanchizadeh, 2022] Geravanchizadeh, Masoud, Parisa Mobasheri, and Hadi Jamshidi Avanaki. "Classification of Iranian Traditional Music Dastgahs Using Features Based on Pitch Frequency." Signal and Data Processing 19.3 (2022): 119-134.

[Farajzadeh, 2023] Farajzadeh, Nacer, Nima Sadeghzadeh, and Mahdi Hashemzadeh. "PMG-Net: Persian music genre classification using deep neural networks." Entertainment Computing 44 (2023): 100518.

[RezezadehAzar, 2018] RezezadehAzar, Shahla, et al. "Instrument-Independent Dastgah Recognition of Iranian Classical Music Using AzarNet." arXiv preprint arXiv:1812.07017 (2018).

[Baba Ali, 2019] Baba Ali, B., A. Gorgan Mohammadi, and A. Faraji Dizaji. "Nava: A Persian Traditional Music Database for the Dastgah and Instrument Recognition Tasks." Advanced Signal Processing 3.2 (2019): 125-134.

[LAYEGH, 2013] LAYEGH, Mahmood ABBASI, Siamak HAGHIPOUR, and Yazdan NAJAFI SAREM. "Classification of the Radif of Mirza Abdollah a canonic repertoire of Persian music using SVM method." Gazi University Journal of Science Part A: Engineering and Innovation 1.4 (2013): 57-66.

[Shirmardi, 2022] Shirmardi, Sahar Sadat, Bagher Babaali and. "Recognition of the Type and Number of Instruments in Iranian Traditional Music." B.Sc Thesis (2022)

[Hemati, 2022] Hemati, Maryam, Bagher Babaali and. "Identify Traditional Musical Instruments with the Help of Machine Learning Methods." B.Sc Thesis (2022)

[Brown, 2022] Brown, Jane A., and Gavin M. Bidelman. "Familiarity of background music modulates the cortical tracking of target speech at the "cocktail party"." Brain Sciences 12.10 (2022): 1320.

[Mohamed, 2022] Mohamed, Abdelrahman, et al. "Self-supervised speech representation learning: A review." IEEE Journal of Selected Topics in Signal Processing (2022).

[Défossez, 2022] Défossez, Alexandre, et al. "High fidelity neural audio compression." arXiv preprint arXiv:2210.13438 (2022).

[Brown, 1991] Brown, Judith C. "Calculation of a constant Q spectral transform." The Journal of the Acoustical Society of America 89.1 (1991): 425-434.