On Self-Supervised Learning in Persian Traditional Music Information Retrieval

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 Wav2Vec 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 Wav2vec 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.

Music can be categorized in many ways, one of which is by its origin. Persian traditional music, a distinct category born from its rich heritage, extends its influence beyond the borders of modern-day Iran. Its captivating melodies and rhythms have resonated with the music of Central Asia, Afghanistan, Pakistan, Azerbaijan, Armenia, Turkey, and Greece [Miller, 2012]. MIR has numerous advantages in the Persian traditional music domain, 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.

Every piece in this dataset has a 44100Hz sampling rate with 16-bit resolution. Nava consists of 7 Dastgahs: Shour, Mahour, Nava, Homayoun, Segah, Chaharghah, and Rast-Panjgah, paired with 5 traditional Persian instruments: Ney, Tar, Sitar, Santoor, and Kamancheh, comprising 1786 solo pieces by 40 Persian artists. Table 1 presents a summary of the Nava dataset, based on the length of available data for each instrument-Dastgah pairing. Nava offers train, test, and evaluation sets, dividing the dataset in a 4:1:1 ratio. 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 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.

	Santoor	Ney	Sitar	Tar	Kamancheh
Shour	3.3	1.2	1.6	1.2	1
Mahour	2..4	1	2.6	1.2	1.1
Chahargah	2.8	1.2	2.6	1.2	1.4
Homayoun	2.8	1.6	1.2	1.2	1.3
Segah	2.5	1.1	1.3	2.1	1.1
Nava	2	1.1	1	1.2	1.1
Rast-Panjgah	1.3	1.2	1.6	1.1	1.2

Table 1: Length of Available Data for Each Instrument-Dastgah Pair Per Hours.

2.2 MIR Approaches in Persian Traditional Music

There have been limited studies on MIR in Persian traditional music, 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, and their perfomances on the downstream tasks testifies to this assertion.

3 Models

3.1 Wav2Vec 2.0

Wav2vec is pretrained on speech data, based on a contrastive loss. Firstly, utilizing a convolutional neural network, each raw input is encoded into a sequence of feature vectors. Secondly, the latent features are quantized using a learned discrete speech units. Finally, masked latent vectors are fed to the model, and the model is encouraged to have the final projections be similar to positive targets and discouraged from being similar to negative ones.

3.2 HuBERT

To address variable lengths for each sound segment and lexicon of audio units for speech representation learning, HuBERT is introduced. Leveraging an offline clustering approach to provide a target label for a BERT-like learning approach, HuBERT was able to outperform Wave2Vec 2.0 in various benchmarks.

3.3 Music2Vec

Music2vec is based on Data2vec 1.0's [Baevski, 2022] multi-modal framework with an aim to predict latent representations of the teacher model. The exponential moving average (EMA) is used to update the student model's weights, as it shares the same architecture with the teacher. Masked input is fed to the student model and is encouraged to predict the average pooling of the teacher model's architecture.

3.4 MusHuBert

Building on HuBERT, MusicHuBERT generates pseudo-labels for masked audio using the K-means algorithm trained on Mel-frequency cepstral coefficients (MFCC) features. In simpler terms, discrete targets are derived from the K-means embeddings of the masked section.

3.5 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.

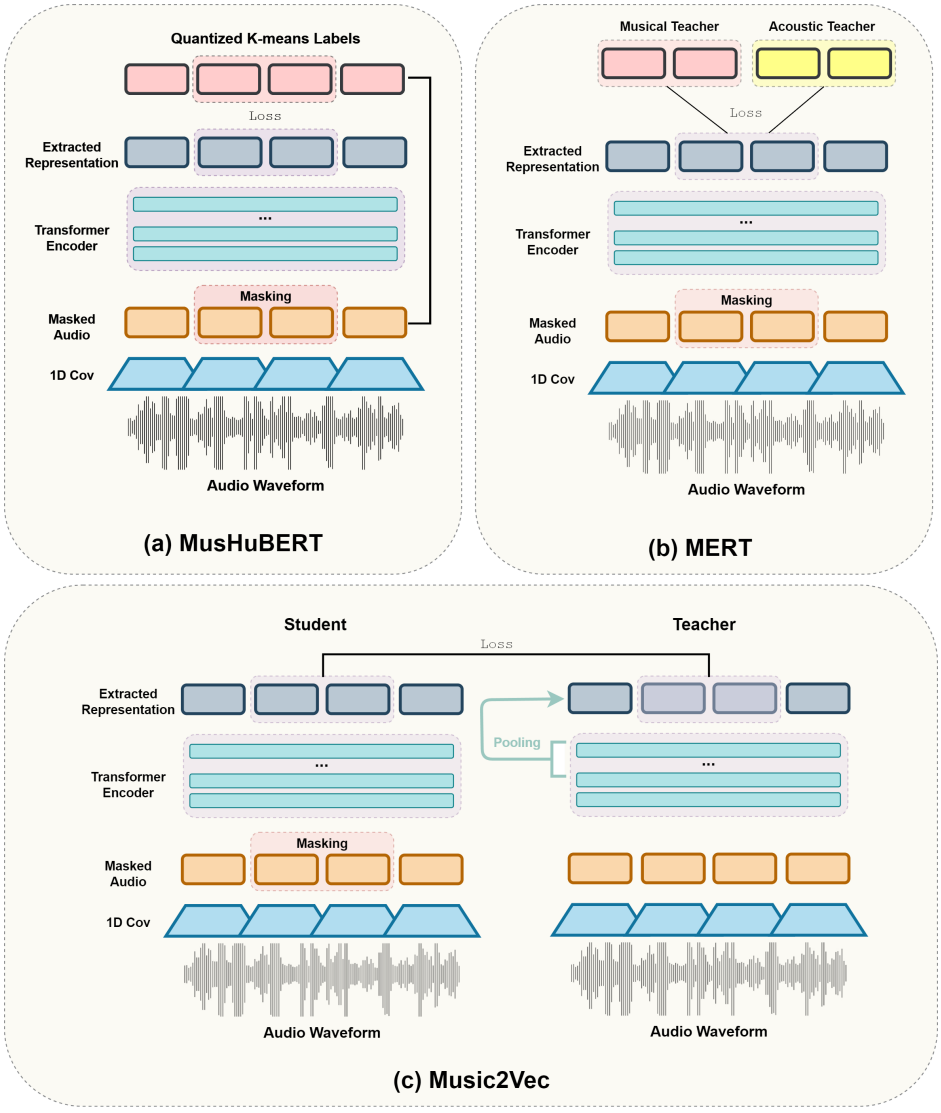

Figure 1: Illustration of the Music2Vec, MusHUBERT, and MERT Pre-training Framework.

4 Experiments and Results

To assess the performance of the five aforementioned models in Persian traditional music, we conducted a series of experiments. Firstly, we examined the impact of audio length on precision by having these models perform in downstream tasks related to instrument and Dastgah recognition, using audio samples of varying durations. Subsequently, we investigated the influence of the transformer's representation layer on accuracy. Following that, we explored the results obtained by employing different Fully Connected (FC) network structures. Lastly, we reported the effects of adjusting the learning rate on model performance in downstream tasks. This section concludes with a comparison of our results with findings from other papers on the Nava dataset.

4.1 Audio Length Effect

In this section, we investigated the impact of audio length on the performance of models. Specifically, we conducted experiments utilizing audio lengths of 3 and 5 seconds, across the Dastgah instrument and artist tasks. To ensure consistency in our experiments, we partitioned the audio pieces into respective 3-second or 5-second segments to obtain our training, validation, and testing datasets. This resulted in the training, validation, and testing datasets being 5/3 times larger in the 3-second tests compared to the 5-second tests. Each experiment was run for 50 epochs.

The findings illustrates that, excpet for the 3-second Dastgah classification task, MERT outpreformed other models. Consequently, we decided to adopt MERT for subsequent experiments because of its promising results in the downstream tasks. Furthermore, it's noteworthy that despite the larger training data size in the 3-second tasks, the models exhibited better performance on the 5-second data, suggesting that longer audio clips might offer certain advantages in our classification tasks. Table 2 summerizes the performances of the models on the defined tasks for 3 and 5 seconds audio lengths.

4.2 Representation Layers Effect

In this subsection, we explored the impact of utilizing various representation layers of a transformer model on the performance of downstream tasks. Our selected models comprised 13 distinct representation layers, each with the potential to impact task accuracy when chosen carefuly. Through experiments, we sought to find the optimal layers whose utilization would yield better performance in the downstream tasks.

	5-second			3-second		
	Instrument	Dastgah	Artist	Instrument	Dastgah	Artist
Music2Vec	78.51	12.79	32.82	74.47	13.47	30.66
MusHuBERT	96.82	18.41	58.80	75.48	**18.66**	40.56
MERT	**97.52**	**19.49**	**63.76**	**96.86**	17.99	**60.53**

Table 2: Comparison of Dastgah, Instrument, and Artist Classification Results for 5-second and 3-second Pieces Across Different Models (The best performance is highlighted in bold).

Overall, our findings indicate that the median layers, specifically the 6th, 7th, and 8th layers, demonstrated superior performance across the downstream tasks. This finding underscores the importance of these intermediate layers in extracting meaningful representations that facilitate task completionHowever, it's noteworthy that the task of Dastgah classification presented unique challenges, as the best performance was achieved at the last layer of the transformer model. This suggests that Dastgah classification demands a more intricate representation to attain a better accuracy. On the other hand, for the artist recognition task, optimal performance was achieved at the 3rd layer, suggesting that this task may require less intricate representations to achieve compelling results. The detailed results of these experiments can be found in Table 4.

4.3 Fully-Connected Architecture Effect

In this subsection, we examined the impact of the fully connected network architecture by testing configurations ranging from 1 hidden layer to 5 hidden layers. It is revealed that the number of hidden layers did not significantly affect the performance of instrument classification tasks, as all architectures consistently achieved compelling results. However, in the context of artist classification, we observed a notable distinction in performance based on the number of hidden layers employed. Specifically, the experiment showed that employing 3 hidden layers outperformed architectures with 5 hidden layers by approximately 5 percent. Moreover, it's noteworthy that across all experiments, excluding configurations with only 1 hidden layer, the fully connected network demonstrated a consistent performance in Dastgah recognition. The detailed results for this experiment is summerized in Table 4.

	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th	13th
Instrument	93.66	95.71	96.39	96.87	96.92	<u>97.11</u>	97.04	**97.47**	97.04	96.99	96.24	96.94	96.36
Dastgah	15.64	17.76	18.77	19.23	18.24	<u>19.01</u>	18.72	17.81	18.46	16.99	16.89	18.70	**19.40**
Artist	66.94	67.71	**69.54**	68.14	66.70	67.25	<u>68.22</u>	67.45	67.86	66.10	64.65	64.48	64.63

Table 3: Performance Comparison of MERT Representation Layers on Downstream Tasks (The best performance is highlighted in bold and the second best is underlined)

	1-layer	2-layer	3-layer	4-layer	5-layer
Instrument	97.59	98.6	98.7	**98.92**	98.65
Dastgah	19.52	**21.54**	20.43	20.94	21.3
Artist	70.07	71.95	**72.43**	68.55	67.75

Table 4: Performance Comparison of Fully-Connected Architecture on Downstream Tasks (The best performance is highlighted in bold).

4.4 Different Learning Rate Effect

4.5 Visualization

4.6 Comparing with Previous Studies

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.

[Miller, 2012] Miller, Lloyd. Music and Song in Persia (RLE Iran B): The Art of Avaz. Routledge, 2012.

[Baevski, 2022] Baevski, Alexei, et al. "Data2vec: A general framework for self-supervised learning in speech, vision and language." International Conference on Machine Learning. PMLR, 2022.