AUDIO QUALITY ASSESSMENT OF VINYL MUSIC COLLECTIONS USING SELF-SUPERVISED LEARNING

Alessandro Ragano[1], *Emmanouil Benetos*[2], *Andrew Hines*[1]

[1]School of CS, University College Dublin, Ireland
[2]School of EECS, Queen Mary University of London, UK

ABSTRACT

Metadata such as mean opinion score (MOS) quality ratings are critical to improve the usability and accessibility of music archive collections. Developing a non-intrusive objective quality metric that predicts MOS of archive music collections is challenging, since it requires labeling large datasets made of real-world recordings, which currently do not exist for this task. In this paper, we show that the self-supervised learning (SSL) model wav2vec 2.0 can be successfully used to predict the perceived audio quality of archive music collections. Using vinyl recordings, we evaluated wav2vec 2.0 on a new dataset of 620 tracks labeled with crowdsourcing. The proposed model shows superior performance to perceptual measures adapted from speech quality prediction. Finally, we propose a new evaluation metric called pairwise ranking accuracy (PRA) that takes into account subjective rater uncertainty by measuring the ability of an objective metric to rank pairs with high-confidence labels.

Index Terms— Perceptual measures of audio quality; objective and subjective quality assessment; self-supervised learning

1. INTRODUCTION

Digital audio archives are provided with metadata to improve user experience and usability. Archive metadata can be manually or computationally created and might include the composer, the carrier, the number of channels, the record label, the year, the genre, etc. The multitude of audio formats and the presence of heterogeneous content have encouraged researchers to develop new computational approaches to improve the accessibility and usability of audio archives. For example, music information retrieval (MIR) tasks such as instrument classification and ethnic group classification were used for non-Western music collections [1, 2] or to analyze and explore large corpora for world music [3] while spoken language technology (SLT) tasks such as automatic speech recognition and speaker identification were used for speech archives [4].

In our previous work [5], we presented the quality of experience (QoE) framework for the evaluation of audio archives. The QoE framework aims to encourage researchers to use a more user-centric automatic approach to evaluate the audio quality of audio archives. In particular, we showed that subjective quality scores are potential useful metadata for digital audio archives, e.g., retrieving best-quality items from archives or detecting the best-quality version of the same composition, which is a typical scenario of classical and jazz collections. Quality score metadata are not provided regularly or are created subjectively by organizations. For example, the Library of Congress described the sound quality of some records in the metadata using the attributes "good" or "bad"[1], which have a very broad meaning. Developing new objective quality metrics would enable automated quality metadata labeling by taking into account user QoE [5].

Predicting the audio quality of archive music collections is not a simple task due to several challenges that we identified [5]. Quality must be predicted with non-intrusive methods, since the reference signal is not available. Large datasets of real-world recordings should be annotated with quality scores, which is time-consuming and expensive. Several music archives include unique recordings such as non-Western cultures or early folk recordings, which makes the creation of a quality metric more difficult due to the low-resource settings. These challenges call for methods that can perform well in limited-annotation scenarios and real-world recordings.

In this paper, we present an objective quality metric for music vinyl collections based on the self-supervised learning (SSL) model wav2vec 2.0 [6]. We focus our work only on vinyl collections, but the results presented can be easily extended to other archive collections. To evaluate the proposed metric, we also contribute with: 1) a dataset of real-world vinyl recordings of Western music annotated with quality scores through crowdsourcing, and 2) a new evaluation performance metric that overcomes some limitations of the correlation coefficients and mean squared error-based metrics typically used for evaluating objective quality metrics for speech data.

The use of wav2vec 2.0 for music quality prediction is motivated by our previous study [7], in which we showed that wav2vec 2.0 can learn general-purpose music representations. Adapted from speech processing, wav2vec 2.0 pre-trained on musical signals turned out to be competitive in instrument classification and pitch classification. The problem of quality prediction in archive audio suffers from the lack of annotated data, and SSL models have been proven to be very effective with only a few minutes of labeled audio for several speech processing tasks [8], speech quality assessment [9, 10], and for music representations [11, 12, 13]. Predicting audio quality requires designing time-consuming listening tests, and labeling large datasets is problematic. By using SSL models, we can learn meaningful representations using a larger unlabeled dataset and then finetune the network with a much smaller labeled data set. The proposed quality prediction models and the dataset used in this work are available on GitHub[2].

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number 17/RC-PhD/3483 and 17/RC/2289_P2. This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

[1]Quality metadata example: https://www.loc.gov/item/2016655162
[2]https://github.com/alessandroragano/music-archive-quality-prediction.git

2. DATASET

The dataset was created by sourcing data from the Boston Public Library Vinyl LP collection [14] and the Vinyl Box collection [15]. These two collections mostly include Western music with different styles (classical, jazz, pop, disco, and electronic). We labeled the quality of the recordings using the absolute category rating scale (ACR) and the Amazon Mechanical Turk (AMT) platform [3].

The preparation of the stimuli was carried out using the same approach that we proposed for real-world speech recordings [16] needed to control the bias that can be generated by random selection of stimuli under uncontrolled conditions. The main idea of our approach is based on creating sessions using stratified random sampling from clusters. To create clusters, we collected 1078 tracks from the two above-mentioned collections and extract 10 seconds from the middle of each track, taking 5 seconds before the middle point of the waveform and 5 seconds after. Following our previous work [16], clustering is performed on 253 audio features, which are obtained by calculating both the actual values and the first-order difference. In this study, we found that K-Means produced better quality clusters on vinyl records compared to HDBSCAN, which was used instead for speech recordings [16]. Sampling the same number of stimuli per cluster can be done only if clusters have the same size. So, we reduced each cluster to the size of the smallest cluster which is 124. This led us to reduce the number of tacks from 1078 to 620 and having 124 tracks per each cluster. We first conducted a pilot test where the feedback collected informed us that 20 stimuli did not affect the fatigue of the participants, which can be explained by the fact that rating on the ACR scale is a simple task. Therefore, each AMT rating session is made up of 4 stimuli per each K-Means cluster, with a total of 20 stimuli. Before the rating session, participants familiarized with the task in a training session which consisted of 12 stimuli sampled with the same cluser-based approach of the rating session.

The listening test followed the ITU P.808 standard for crowdsourcing speech quality evaluation to create trapping questions, check the use of 2 channels, ask participants about their hearing ability, and ask with which device they performed the test [17]. The tracks are converted to a lossy format with high-efficiency advanced audio coding (HE-AAC) at 320 kbps, which avoids the potential stalling that can be caused by network problems of the participants while still preserving audio quality. Loudness normalization using EBU R 128 [18] is applied to all stimuli to avoid that the quality is biased by loudness. Before the training and the rating sessions, participants performed a setup session where they could adjust the device volume and they were asked to add 2 or 3 digits that are played only in the left or right channels in order to check for a functioning stereo configuration.

Each participant was paid 0.50¢ per rating session and a bonus of 0.10¢ has been assigned to participants who completed more than 15 sessions. To reduce participant fatigue, no more than 20 sessions were allowed for the same recruiter. The trapping questions have been used to detect unreliable participants or potential cheaters. The trapping stimulus begins with music followed by a message that says "This is an interruption, please select the answer x" where x is one of the 5 categories on the ACR scale (bad, poor, fair, good, and excellent). 60 trapping questions have been created using 12 tracks that were not among the rating stimuli and 5 messages, one for each category of the ACR scale. Trapping questions were randomly distributed throughout the sessions. Participants who did not meet at least one of the following conditions were excluded from the response analysis: answering incorrectly the math question, answering incorrectly the trapping question, declaring not having a normal hearing ability, declaring not having headphones available, if their score variance was lower than 0.1.

[3] University College Dublin approved this study as a low-risk study

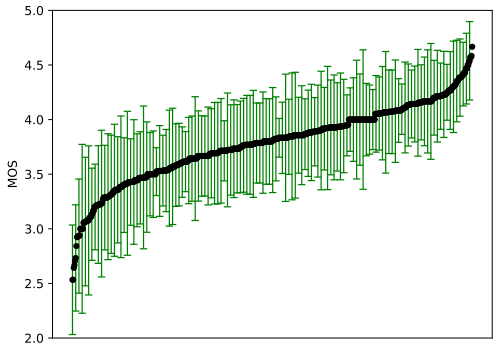


Fig. 1: MOS of 620 vinyl recordings sorted from lowest to highest, with 95% confidence intervals shown every 5 tracks.

A total of 506 participants and 822 sessions were collected with 469 sessions that were marked as valid. The valid number of participants per track ranges from a minimum of 10 to a maximum of 21 participants with a mean of $\approx$ 15 participants per track. For each track, we compute the mean opinion score (MOS), which is shown in Figure 1. The prepared dataset is called Vinylset.

3. MODEL

The proposed objective quality metric is based on fine-tuning wav2vec 2.0, which is a contrastive learning-based approach where the model learns to distinguish a target sample (positive) from distractors (negative) using a convolutional feature encoder followed by a context network based on the Transformer architecture [6]. To use wav2vec 2.0 with music, we pre-trained the architecture using MusicNet [19] for 1790 epochs using the repository made available on fairseq [20]. Following the instructions given in the repo, the MusicNet recordings have been split into 20-second samples. To increase the dataset size we used an hop size of 10 seconds. Furthermore, we downsampled the files to 16 kHz, which is the expected sampling rate for wav2vec 2.0. The model was trained on a NVIDIA A100 40 GB GPU and took 7 days to finish. Fine-tuning of wav2vec 2.0 on the Vinylset corpus is performed by taking the mean of the features of the last Transformer block to remove the time dimension. A linear layer is used to predict MOS scores.

4. EXPERIMENTAL DESIGN

4.1. Cross-Validation

Since Vinylset includes 620 observations, using only one split into training, validation, and test set could generate biased results on the particular partition. For this reason, this experiment proposes to use stratified k-fold cross-validation. The number of folds must meet the criteria that the MOS distribution should be similar in training, validation, and test sets. A high number of folds helps to reduce the variance of the performances since the model is trained on more training partitions. However, setting k too high introduces some disadvantages. For example, there is a higher chance that the MOS distributions of the validation and test sets are too dissimilar between the different folds and that stratification cannot be achieved successfully.

By visually inspecting the MOS distributions of the training, validation, and test sets in the folds, and by dividing the MOS range into 15 classes, we found that $k = 3$ is an appropriate value for this dataset. Indeed, using a higher number of folds gave partitions that were too dissimilar from each other, in particular, at the extreme MOS values. Using 3-fold stratified cross-validation and MOS classes 15, each fold is divided into ≈67%, ≈16%, ≈16% for training, validation, and test sets, respectively.

4.2. Baselines

No baseline can be found to predict the quality of archive music collections. For this reason, we decided to compare the proposed model against non-intrusive deep learning models developed for speech quality prediction as shown in Table 1.

4.2.1. Random Labels

One of the baseline models consists of replacing the real Vinylset labels with random labels. The random label model is used to understand the reliability of the collected labels. Random labels are generated by sampling from a Gaussian distribution with mean and standard deviation calculated from the real Vinylset labels. Sampling is performed before training, and labels are fixed during training.

4.2.2. NISQA

The NISQA metric was originally designed for super-wideband speech quality prediction and consists of three main blocks: a framewise ConvNet, a self-attention network to model the time dependency, and an attention-pooling network to predict MOS [21]. We trained two different versions of NISQA. A model that uses all the default settings of the NISQA repository and a second model that uses the L1 loss instead of the L2 loss for both optimization and early stopping. Since we used the L1 loss in all other models, training NISQA with the same loss function of the proposed model gives us a fairer comparison.

4.2.3. Pre-Trained Models

In our previous work [22] we showed that pre-training a ConvNet from a degradation classifier and from deep convolutional embedded clustering (DCEC) improves speech quality prediction in the limited-annotation scenario. For training these models and achieving a fair comparison with NISQA we use a simplified version of NISQA that we call ConvMaxPool. We take the same framewise ConvNet and replace the self-attention network with a temporal max-pooling layer and the attention-pooling network with a linear layer. By applying these changes, we ensure that the main contribution to model performance is given by the features learned in the ConvNet and not by advanced techniques such as the self-attention network of the original NISQA model. To train the degradation classifier, we create a synthetic dataset with the following degradations: clip, codecs, background noise, reverberation, and echo. The model is trained to classify six classes, i.e. five degradations plus the clean signal. 10,000 samples are randomly taken from the Free Music Archive (FMA) dataset [23] and every track is degraded with the five degradations, collecting 60,000 samples in total. The model pre-trained with the degradation classifier is called ConvMaxPool Degr. Class. in Table 1. The DCEC model is trained on the same overlapped segments of MusicNet and finetuning is carried out with both single-task and multi-task learning (MTL) as done in [22]. These two models are called ConvMaxPool DCEC and ConvMaxPool DCEC MTL in Table 1.

5. TRAINING

All models are trained to minimize the L1 loss. The proposed model is trained using batch size 4 and optimized with Adam using a learning rate of $1e-5$ for the pre-trained part and $1e-4$ for the linear layer at the output. All ConvMaxPool-based models are trained using the same input features of the NISQA model, which is a log-mel spectrogram calculated with window length 20 ms, hop length of 10 ms, and 48 mel bands. ConvMaxPool-based models are fine-tuned with batch size 16 and optimized with Adam using a learning rate of $1e-4$ for the pre-trained framewise CNN and $1e-3$ for the output linear layer. Training was stopped if the loss function calculated in the validation set did not decrease after 20 epochs. We found that the performance on the validation set increased when using a lower learning rate only in the pre-trained layers.

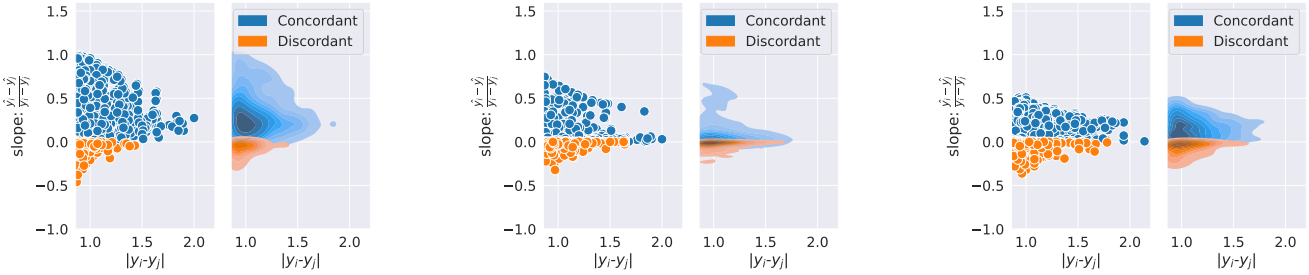
6. RESULTS & DISCUSSION

Evaluating objective quality metrics is typically carried out using the root mean squared error (RMSE), Pearson's correlation coefficient (PCC) and Spearman's rank correlation coefficient (SRCC) calculated per condition [24]. However, since the dataset used is made up of real-world recordings, we must evaluate performance per recording rather than using multiple stimuli with a common condition. The predictions are mapped using a third-order polynomial as recommended in ITU P.1401 [24] that adjusts for subjective test bias.

The results in Table 1 show that w2vMOS outperforms all the baselines in all evaluation criteria. Unlike studies on speech quality prediction, this task shows a relative lower PCC or SRCC. We believe that the lower scores found in our experiments could be due to two reasons. First, we are not aggregating performance by condition, which typically improves correlation scores. In fact, two degraded stimuli created by applying the same degradation condition to two different clean recordings might be labeled with different MOS values. Aggregation of predictions in the performance evaluation cancels out these individual differences and improves the performance scores. Another reason is the meaning of the MOS scores in the proposed corpus Vinylset. Real-world stimuli from vinyl recordings represent a much harder scenario since the degree of acceptability of quality might be high even if the recordings are noisy. Some participants may find some classical or jazz recordings with perceivable hiss pleasant and, therefore, they might rate the quality higher than others. This phenomenon can be observed via the 95% MOS confidence intervals (CI) shown in Figure 1. We can see that several sam-

Table 1: Performance evaluation using RMSE, PCC, and SRCC.

	RMSE	PCC	SRCC
ConvMaxPool	0.32 ± 0.009	0.36 ± 0.058	0.30 ± 0.097
ConvMaxPool Autoencoder	0.32 ± 0.016	0.36 ± 0.107	0.31 ± 0.131
ConvMaxPool Degr. Class. [22]	0.31 ± 0.012	0.40 ± 0.074	0.38 ± 0.090
ConvMaxPool DCEC [22]	0.32 ± 0.015	0.38 ± 0.091	0.34 ± 0.105
ConvMaxPool DCEC MTL [22]	0.31 ± 0.011	0.39 ± 0.064	0.34 ± 0.080
NISQA (L1 loss) [21]	0.33 ± 0.015	0.34 ± 0.035	0.33 ± 0.034
NISQA (default) [21]	0.36 ± 0.020	0.38 ± 0.103	0.36 ± 0.007
w2vMOS Rand. Labels	0.34 ± 0.007	0.19 ± 0.068	0.11 ± 0.067
w2vMOS	$\mathbf{0.29 \pm 0.017}$	$\mathbf{0.50 \pm 0.079}$	$\mathbf{0.47 \pm 0.066}$



(a) w2vMOS - PRA= 89.74%

(b) w2vMOS Rand. Labels - PRA= 66.13%

(c) NISQA (L1 loss) - PRA= 77.95%

Fig. 2: Ratio between slope and ground truth absolute difference of (a) w2vMOS, (b) w2vMOS Random Labels, and (c) NISQA (L1 loss) of the three test partitions. PRA indicates the fraction of concordant pairs (blue) over the total pairs (blue + orange).

ples with close MOS labels show a CI that is close to 1 (the average CI calculated using all 620 tracks is ≈ 0.87). This implies that there is a high degree of uncertainty in the labels collected, especially in the samples whose MOS is close to the mean of the distribution.

Performance scores obtained with RMSE, PCC and SRCC do not consider the uncertainty of the participants propagated in the labels as discussed above. For this reason, we propose a new evaluation metric called Pairwise Ranking Accuracy (PRA). Let $\mathcal{N}$ denote the test set, y_n the ground truth MOS of the n-th observation, $\hat{y}_n$ the predicted MOS of the n-th observation and $\mathcal{S} = \{(i,j)|i,j \in \mathcal{N}, |y_i - y_j| > \tau\}$ the set of all the combinations in the test set subject to the constraint $|y_i - y_j| > \tau$, the PRA is defined as:

$$PRA = \frac{1}{|S|} \sum_{i<j} \frac{\text{sgn}(y_i - y_j)\,\text{sgn}(\hat{y}_i - \hat{y}_j) + 1}{2}, \forall i,j \in \mathcal{S}.$$

PRA measures the ability of an objective quality metric to rank the MOS of pairs whose MOS distance is greater than a threshold τ. The latter is set to $\tau = \frac{1}{|\mathcal{D}|}\sum_{k=1}^{|\mathcal{D}|} CI_k$ where CI_k is the 95% confidence interval of the k-th observation in the training set $\mathcal{D}$. The PRA calculates the number of concordant pairs over the total combinations in the constrained set $\mathcal{S}$. The idea behind the proposed performance measure is that an objective quality metric is robust if it is able to rank stimuli of pairs whose MOS distance have higher confidence. To measure which pairs have high-confidence labels, we take the stimuli where the MOS scores differ at least by the average confidence intervals since it is expected that there is higher chance that the rank of these pairs will not change if we repeat the test with different participants. Note that we did not take just the individual confidence interval of each track in the dataset since they are generated by different groups of listeners exposed to different stimuli. The threshold is calculated in the training set to avoid information leakage from the test set. In practice, it does not make much difference in our dataset since the training and test subsets are two samples of the same distribution. If the test set is sampled from a different distribution, the threshold should be calculated in the test set. Notice that the Kendall's Tau coefficient is a common statistical measure applied to evaluate pair ranking performance. However, Kendall's coefficient evaluates all the possible combinations of pairs while we take a subset of pairs with the constraint $|y_i - y_j| > \tau$.

The results using PRA are shown in Table 2 and indicate the superiority of w2vMOS, which correctly detects the rank of $\approx 90\%$ of the high-confidence pairs. A visualization of PRA is shown in Figure 2. Higher PRA values are obtained if the density of the concordant pairs (blue) increases or if the density of the discordant pairs (orange) decreases. Concordant pairs and discordant pairs of w2vMOS with random labels are shown to be concentrated around slope 0 which means that the pair rank is random. It can be seen that the discordant pairs of the w2vMOS model are closer to the lowest value of the x-axis which corresponds to the threshold τ, indicating that w2vMOS is less confident when ranking stimuli with a closer MOS. Regarding NISQA (L1 loss) and w2vMOS trained with random labels, we can see how the discordant pairs are more distant from the origin compared to w2vMOS, indicating that these two models do not rank correctly the pairs where the MOS distance is very high.

Table 2: Comparison of objective quality metrics using Pairwise Ranking Accuracy (PRA).

	PRA (%)
ConvMaxPool	77.19 ± 4.33
ConvMaxPool Autoencoder	79.99 ± 7.81
ConvMaxPool Degr. Class. [22]	82.72 ± 4.88
ConvMaxPool DCEC [22]	84.85 ± 4.51
ConvMaxPool DCEC MTL [22]	85.32 ± 2.49
NISQA (L1 loss) [21]	77.95 ± 8.77
NISQA (default) [21]	80.8 ± 10.88
w2vMOS Rand. Labels	66.13 ± 8.32
w2vMOS	$\mathbf{89.74 \pm 3.69}$

7. CONCLUSIONS

In this paper, we show that fine-tuning wav2vec 2.0 is a promising solution to estimate the quality of vinyl music collections. The performance of wav2vec 2.0 is superior to objective quality metrics based on supervised learning and deep clustering feature representations. Furthermore, we introduce a new dataset of real-world vinyl recordings labeled with crowdsourcing, and we present the PRA performance metric which takes into account the uncertainty of the participants. In the future, we will understand if the parameters of wav2vec 2.0 (e.g. window length, number of Transformer blocks) can be modified to suit better music signals since the model is originally proposed for speech representations. Also, we will evaluate objective quality metrics for audio codecs on musical signals and for more archive formats such as wax cylinders and shellac discs.

8. REFERENCES

[1] Thomas Lidy, Carlos N Silla Jr, Olmo Cornelis, Fabien Gouyon, Andreas Rauber, Celso AA Kaestner, and Alessandro L Koerich, "On the suitability of state-of-the-art music information retrieval methods for analyzing, categorizing and accessing non-western and ethnic music collections," *Signal Processing*, vol. 90, no. 4, pp. 1032–1048, 2010.

[2] Olmo Cornelis, Micheline Lesaffre, Dirk Moelants, and Marc Leman, "Access to ethnic music: Advances and perspectives in content-based music information retrieval," *Signal Processing*, vol. 90, no. 4, pp. 1008–1031, 2010.

[3] Maria Panteli, Emmanouil Benetos, and Simon Dixon, "A review of manual and computational approaches for the study of world music corpora," *Journal of New Music Research*, vol. 47, no. 2, pp. 176–189, 2018.

[4] Jerry Goldman, Steve Renals, Steven Bird, Franciska De Jong, Marcello Federico, Carl Fleischhauer, Mark Kornbluh, Lori Lamel, Douglas W Oard, Claire Stewart, et al., "Accessing the spoken word," *International Journal on Digital Libraries*, vol. 5, no. 4, pp. 287–298, 2005.

[5] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines, "Automatic quality assessment of digitized and restored sound archives," *Journal of the Audio Engineering Society*, vol. 70, no. 4, pp. 252–270, April 2022.

[6] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[7] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines, "Learning music representations with wav2vec 2.0," *arXiv preprint arXiv:4568234*, 2022.

[8] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al., "Self-supervised speech representation learning: A review," *arXiv preprint arXiv:2205.10643*, 2022.

[9] Helard Becerra, Alessandro Ragano, and Andrew Hines, "Exploring the influence of fine-tuning data on wav2vec 2.0 model for blind speech quality prediction," *Proc. Interspeech*, pp. 4088–4092, 2022.

[10] Wei-Cheng Tseng, Chien-yu Huang, Wei-Tsung Kao, Yist Y Lin, and Hung-yi Lee, "Utilizing self-supervised representations for mos prediction," *Proc. Interspeech*, pp. 2781–2785, 2021.

[11] Ho-Hsiang Wu, Chieh-Chi Kao, Qingming Tang, Ming Sun, Brian McFee, Juan Pablo Bello, and Chao Wang, "Multi-task self-supervised pre-training for music classification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 556–560.

[12] Hongyuan Zhu, Ye Niu, Di Fu, and Hao Wang, "Musicbert: A self-supervised learning of music representation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3955–3963.

[13] Andrew N Carr, Quentin Berthet, Mathieu Blondel, Olivier Teboul, and Neil Zeghidour, "Self-supervised learning of audio representations from permutations with differentiable ranking," *IEEE Signal Processing Letters*, vol. 28, pp. 708–712, 2021.

[14] "The vinyl boston public library collection," `https://archive.org/details/vinyl_bostonpubliclibrary?tab=about`, Accessed: 2022-10-06.

[15] "The vinyl box collection," `https://archive.org/details/the-vinyl-box?tab=about`, Accessed: 2022-10-06.

[16] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines, "Development of a speech quality database under uncontrolled conditions," *Proc. Interspeech 2020*, pp. 4616–4620, 2020.

[17] Babak Naderi and Ross Cutler, "An open source implementation of itu-t recommendation p. 808 with validation," *Proc. Interspeech 2020*, pp. 2862–2866, 2020.

[18] R EBU-Recommendation, "Loudness normalisation and permitted maximum level of audio signals," 2011.

[19] J. Thickstun, Z. Harchaoui, and S.M. Kakade, "Learning features of music from scratch," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[20] Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 48–53.

[21] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech*, 2021, pp. 2127–2131.

[22] Alessandro Ragano, Emmanouil Benetos, and Andrew Hines, "More for less: Non-intrusive speech quality assessment with limited annotations," in *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2021, pp. 103–108.

[23] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "FMA: A dataset for music analysis," in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[24] International Telecommunication Union (ITU), "P.1401 Methods, metrics and procedures for statistical evaluation, qualification and comparison of objective quality prediction models," 2020.