FROM AUDIO ENCODERS TO PIANO JUDGES: BENCHMARKING PERFORMANCE UNDERSTANDING FOR SOLO PIANO

Huan Zhang[1]   **Jinhua Liang**[1]   **Simon Dixon**[1]

[1] Center for Digital Music, Queen Mary University of London, UK

`huan.zhang@qmul.ac.uk`

ABSTRACT

Our study investigates an approach for understanding musical performances through the lens of audio encoding models, focusing on the domain of solo Western classical piano music. Compared to composition-level attribute understanding such as key or genre, we identify a knowledge gap in performance-level music understanding, and address three critical tasks: expertise ranking, difficulty estimation, and piano technique detection, introducing a comprehensive Pianism-Labelling Dataset (PLD) for this purpose. We leverage pre-trained audio encoders, specifically Jukebox, Audio-MAE, MERT, and DAC, demonstrating varied capabilities in tackling downstream tasks, to explore whether domain-specific fine-tuning enhances capability in capturing performance nuances. Our best approach achieved 93.6% accuracy in expertise ranking, 33.7% in difficulty estimation, and 46.7% in technique detection, with Audio-MAE as the overall most effective encoder. Finally, we conducted a case study on Chopin Piano Competition data using trained models for expertise ranking, which highlights the challenge of accurately assessing top-tier performances.

1. INTRODUCTION

Traditional music understanding tasks focus on composition-level attributes: key, tempo, genre and instrumentation are widely explored [1–3]. These attributes are not only tagged individually via end-to-end approaches but have also been the focus of foundation models and various musical representations aimed at learning them in a unified manner [4, 5], facilitating cross-modal understanding [6, 7].

However, a large portion of human music activity is focused not on the composed songs or pieces themselves, but on the process of learning and performing them [8]. Despite its great importance to the vast community of students, teachers and musicians, the ability to understand performance nuances (challenging techniques, skill varieties, stylistic differences, difficulty grading, etc.) has not been grasped by machines. Sporadic experiments [9] of these tasks are conducted, often on small-scale [10, 11] or proprietary [12–14] datasets. Performance understanding, in contrast to the more recognized composition-level music understanding, suffers from scarcity of data [15, 16], ambiguity of tasks [17], and the inherent complexity of modelling and representing expressive elements in performances [18, 19].

Meanwhile, unified representations and foundation models have advanced several fields by providing robust and versatile frameworks [6, 20], demonstrating their potential to overcome challenges related to data scarcity and task specificity. Building on this precedent and their applications to compositional-level understanding [21], we extend the capacities of pre-trained audio encoders such as MERT [22] and MULE [5] into the performance-understanding realm, investigating the shared knowledge between composition- and performance-level understanding: Do pre-trained audio encoders capture performance nuances? Can they categorize performance-related attributes? If not, how can we improve their performance?

This work is a first step in filling the gaps within the performance understanding realm. Applying domain-adaptation to pre-trained audio encoders, we work towards a *piano judge* that specializes in ranking performers' skill level, determining the given repertoire's difficulty and core techniques, thus pursuing a human piano teacher's capability and paving the way to performance understanding in an educational context. Our contributions [1] include:

1. We benchmark three tasks in the realm of audio performance understanding: expertise ranking, difficulty estimation, and solo piano technique detection.
2. We leverage four audio representation learners (Jukebox, Audio-MAE, MERT, DAC) and compare their capabilities in tackling the downstream tasks.
3. We release Pianism-Labelling Datasets (PLD) with detailed labeling curated for the three tasks, the first large-scale dataset (136 hrs in total) that aims to address performance understanding.
4. We fine-tune DAC [23] and AudioMAE [24] by domain-adaptation with solo piano, and compare their performances with pre-trained versions.
5. We conduct a case study on Chopin Piano Competition data (*ICPC-2015*), exploring how a trained expertise ranking model can be transferred to rank candidates in the most prestigious competition of the pianistic scene.

---

[1] Code available at: https://github.com/anusfoil/PianoJudges

2. RELATED WORK

2.1 Performance and education focused understanding

The exploration of performance through recordings provides rich resources for music understanding. Automatic performance analysis (APA) [17] delves into dimensions from dynamics [25, 26] to timing [27, 28], forming the basis for tasks such as performer identification and automatic music performance assessment (MPA) [15]. The former seeks to attribute performances to their respective musicians based on stylistic and technical signatures [29, 30], and the latter aims to evaluate the quality and expression of performances. MPA approaches can be further divided by the level of proficiency. For novice players, the emphasis is on technical accuracy, ensuring correct notes and rhythm via score alignment [31, 32] or detecting conspicuous mistake regions [14] in a score-free context. Advanced performers are assessed by their expression and musicality, usually in the form of predicting rating scores on multiple dimensions [12, 33]. Recently, the release of feedback-based assessment data [34, 35] offers the possibility to conduct multimodal MPA in a more personalized manner.

On the other hand, the analysis of performances in an educational context emphasizes the identification of challenges and learning opportunities within the repertoire: expert-annotated difficulty level is predicted from symbolic scores [36] via a machine learning classification approach that merges musicologically-inspired score features. At a more granular level, we would like to identify instrument-specific techniques that demand practice. For example, techniques such as *acciacatura* and *portamento* on Chinese bamboo flute can be identified from spectro-temporal patterns [37], but similar problems have yet to be explored on piano because of the homogeneity of piano sound. Other learning aid information such as fingering [9, 38] and bowing [39] can also be predicted. In this work, we focus on expertise, difficulty and technique estimation by extracting relevant information from performance audio. We are the first of its kind in piano technique detection and expertise ranking, and an audio representation approach in compare with the current difficulty estimation [9, 40].

2.2 Leveraging audio representations for downstream tasks

The surge of learning audio representations was originally motivated by the generative models such as AudioLM [41] and MusicLM [42]. Jukebox [43] is a generative model trained on 1.2M songs. Subsequent work [4, 44] has shown that Jukebox's representations can be effective features for task-specific linear classifiers. Jukebox embeddings have also been employed in multimodal learning [6] of music captioning and reasoning tasks. MERT [22] uses masked language modelling (MLM) style acoustic self-supervised pre-training. With a music teacher and an acoustic teacher, MERT demonstrates good performance in downstream music understanding tasks and extends its music understanding ability into question answering and captioning [45] by generating music representations to aid language model.

Audio-MAE [24] is a vanilla 12-layer transformer that learns to reconstruct randomly-masked spectrogram patches. The output feature map from the penultimate block of an Audio-MAE encoder has been used to encode fine-grained patterns in audio [46]. Different from previous approaches, Descript-Audio-Codec (DAC) [23] is a neural audio compression autoencoder that compresses high-dimensional signals into lower dimensional discrete tokens. DAC has been proven useful in a generative context [47], but there have been few attempts to explore it with downstream understanding tasks [48].

The four aforementioned audio representations are chosen for our investigation. Since they are constructed from different theoretical approaches (quantized codecs vs. continuous spectrograms) and trained on different data (general audio vs. music), this variety presents an opportunity to evaluate the extent to which the encoded information contributes to performance understanding.

3. METHODOLOGY

3.1 Downstream problem definitions

3.1.1 Expertise ranking

We formulate our assessment into a ranking problem: given audio performances p_1 and p_2, which one has the higher expertise? We define three coarse levels of expertise (beginner, advanced and virtuoso), represented by integers 0, 1 and 2, respectively, and define a function Q which maps a performance to one of these levels. Instead of directly predicting the absolute expertise level Q, we learn a 2-way or 4-way ranking function between each pair of performances from different levels, R_2 or R_4, as below:

$$R_2 = \begin{cases} 0 & Q(p_1) < Q(p_2) \\ 1 & Q(p_1) > Q(p_2) \end{cases} \qquad R_4 = \begin{cases} 0 & Q(p_2) - Q(p_1) = 2 \\ 1 & Q(p_2) - Q(p_1) = 1 \\ 2 & Q(p_1) - Q(p_2) = 1 \\ 3 & Q(p_1) - Q(p_2) = 2 \end{cases} \quad (1)$$

The motivation behind is to teach the model a relative notion of expertise, instead of an absolute level or category of the performance quality. In real life and competition settings (as will be discussed in Sec 5.1.1), we are more interested in the comparative skill level among a set of candidates.

3.1.2 Difficulty estimation

Following the literature [9, 18, 36] on difficulty level prediction, we formulate the problem as a classification task with 9 difficulty classes, given the dataset described in Section 4.2, which has 9 levels of difficulty annotation. Given that the difficulty annotation is subjective and boundaries between levels are fuzzy, we also report the results of 3-class estimation by merging the level groups, as in [36].

3.1.3 Technique identification

Given a piece, a piano teacher can immediately identify the most challenging passage(s) that would require students hours of practice to master: intense octave runs, fast flowing scales, repeating notes that require finger iteration, etc.

Encoder	C	F (Hz)	Dim
Jukebox	2048	345	64
MERT	-	75	1024
Audio-MAE	-	51.2	768
DAC	9×1024	87	1024
Spectrogram	-	150	128

Table 1. Specifications of the audio encoders as well as the spectrogram baseline: C is the codebook size, F is the frame rate in Hz and Dim is the hidden dimension of the embedding (mel-bins for spectrogram).
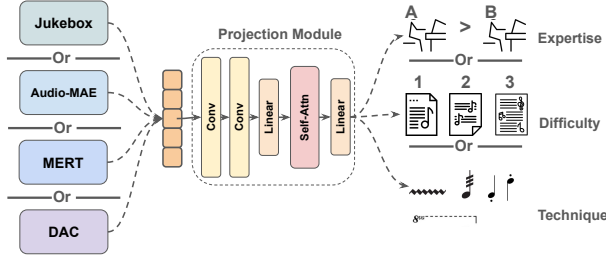

Figure 1. Overview of our tasks and experiment pipeline. For the expertise ranking, two audio embeddings are concatenated in the time dimension.

In the technique-specific dataset (Section 4.3), we include 7 common techniques and formulate a multi-label classification task for technique identification. Given that our labels are relatively sparse, we also experiment with the case of single-label prediction in which predicting any one of the multiple labels is considered correct.

3.2 Audio embeddings and encoder fine-tuning

An overview of the used audio embeddings is given in Table 1. For Jukebox, we employ the 345 Hz sample rate encoding, and for Audio-MAE, the 768-dimensional embedding is taken from the ViT-B Transformer encoder. Additionally, we considered a spectrogram baseline, as a low-level representation to compare with the trained embeddings. We use 128 mel bins, an FFT of 400 samples, and hop size of 160 samples, resulting in a spectrogram with frame rate of 150 Hz and 128 dimensions that feeds into the prediction head module like the trained embeddings.

We examine whether fine-tuning the two generic-audio-trained encoders DAC and Audio-MAE with domain-specific data results in a performance boost . The two encoders are fine-tuned using their original self-supervision objective on around 2k hours of solo piano recordings, from datasets of MAESTRO [49], ATEPP [50], SMD [51], Mazurkas [2] as well as the novel PLD data introduced in this work. For DAC, the fine-tuning lasts for 25k iterations while the Audio-MAE is fine-tuned for 64 epochs.

3.3 Experiments

For all encoders, we first compute 10-second segment audio embeddings (or spectrograms), and include a maximum of 5

---
[2] http://www.charm.rhul.ac.uk/index.html

	task type	classes	tracks	len. (s)
Expertise	Multi-class	2 or 4	1694	167.4
ICPC-2015	Multi-class	2	137	1827.0
Difficulty	Multi-class	3 or 9	737	269.8
Techniques	Multi-label	7	222	45.5

Table 2. Dataset statistics (number of classes, number of tracks and average duration) for each task.

minutes (30 segments) of audio as input with padding. The concatenated embedding of each audio track is of shape $(30, F \times 10, D)$ where F is the frame rate and D is the embedding dimension as shown in Table 1.

Given the audio embedding, we transform it through a prediction head module that consists of two 2D convolutional layers ($n_{kernel} = 7, n_{stride} = 5$), one linear layer to align different input dimensions, and one self-attention layer ($n_{heads} = 2, d = 128$), followed by a final linear layer that projects to the desired classes of each task. A full pipeline of the experiments is shown in Figure 1. As is standard practice [4, 44], we maintain a straightforward projection module design, aiming to minimize its influence on the probing performance.

Regarding each individual task, we run a grid search on the hyper-parameters for learning rate, weight decay, batch size, etc. The details for the final training parameters for each task are documented in the project page [3] . All fine-tuning and training are conducted on one NVIDIA A5000.

4. PIANISM-LABELING DATASET

The pianism-labeling dataset (PLD) includes audio and annotations for three notions that are centrally relevant to pianism and piano education: **expertise**, piece **technique** and **difficulty**, where dataset statistics are specified in Table 2. All of the labeling, metadata correspondence, as well as examples are available on the project page.

4.1 Expertise

We curated a collection of solo piano recordings from YouTube, each annotated with an expertise level. Their categorization was based on information gleaned from the YouTube channels' descriptions, which provided insights into the background of the recordings. This categorization process was validated by two college-level piano students to ensure accuracy.

- **Beginner** (562): Amateur level, featuring mostly adult self-taught learners' practice recordings.
- **Advanced** (570): Performances of music students and junior competition recordings.
- **Virtuoso** (562): Famous pianists' recordings sourced from the ATEPP [50] dataset. To balance with the other groups, we randomly select a subset of 562 of the 11K recordings.

The repertoire of selected performances is mainly focused on the Western Classical repertoire, with some rearranged folk and pop songs at the Beginner level. Indeed,

---
[3] https://bit.ly/3SYzozY

it is challenging to align the performed repertoire across levels since the complexity of played pieces increases with the expertise: e.g. Beginners' pieces are shorter (av. 128.9 s) than Advanced (201.1 s) or Virtuoso (171.8 s) tracks.

In experiments, the three levels are first individually split into train and test subsets and then paired up randomly. Each recording only shows up once in the pairs to prevent leakage, which results in 2694 pairs in training.

4.1.1 ICPC-2015

In this task we aim to assess whether the learnt comparative ranking objective can be applied to the professional domain, the International Chopin Piano Competition, with data gathered from the 2015 edition (*ICPC-2015 dataset*)[4]. We employ only the preliminary round performances to ensure limits on the length and instrumentation (i.e. solo), and assume that the overall better players clearly demonstrate their skills in the preliminary round performance. Out of 160 candidates, 137 recordings are successfully retrieved.

We compile the data into ranking pairs similar to Section 4.1, by first assigning a score $S(c)$ for each candidate based on their progression into the following rounds. For every round that the candidate passes into, the score is incremented with 1 point. For candidates a, b with their respective scores $S(a), S(b)$, all preliminary round recordings are formed into pairs with ranking as in Eq. 1. As shown in Table 2, the preliminary round recordings have an average duration of 30 minutes. Thus, we obtain paired ranking results for each pair of 5-minute segments (30 segments in total) and use majority voting to obtain the final rank among two recordings.

4.2 Difficulty

We employ the *Can I Play It?* (CIPI) [36] dataset for our task of difficulty prediction. Given that the original dataset is sourced in symbolic MusicXML, we obtain the performance audio from YouTube by querying the metadata followed by manual correction to enforce piece alignment. Note that the performances are sourced from different levels of playing rather than virtuoso recordings only, with the aim of learning a more general view of audio difficulty. In the CIPI dataset, difficulty labels are annotated by Henle Verlag[5], a renowned publisher in the music education community. The ratings range from 1-9 and span 29 composers. Note that we split the movements from sonata or other multi-movement compositions, resulting in the 737 audio tracks shown in Table 2 compared to 637 compositions in the original metadata. We also use the same train-test split as the original dataset.

4.3 Techniques

The technique dataset contains 222 recordings with an average duration of 45 seconds, demonstrating one or more canonical piano techniques from seven categories taken from piano practice literature [52]. The excerpts are taken

from etude books like Beyer or Czerney, or passages from performance repertoire (e.g. dense octave run from *Chopin op.25 no.10*). Besides YouTube sourcing, 41 out of 222 recordings are recorded by the authors, if the specific passages containing the techniques are not publicly available in any recording. The categories of techniques are:

- **Scales** (48): Pure scale run across octaves. Can be both hands or one hand.
- **Arpeggios** (40): Pure arpeggio run across octaves, or music passages that are accompanied with arpeggios, or melody that is constructed on arpeggiated chords.
- **Ornaments** (31): Including grace notes, trills, *mordents*, *acciacatura*. Note that we do not balance these subclasses, and the most common ornament in our samples is grace note.
- **Repeated notes** (35): Musical passages that feature a series of repeated single notes.
- **Double notes** (36): Musical passages that feature sequences of simultaneous intervals (mostly thirds, but also fourths and sixths), where the intervals are performed with one hand.
- **Octave** (35): We differentiate octaves from double notes because of their sheer importance in piano repertoire, as well as its distinctive sonority.
- **Staccato** (41): Musical passages that are predominately performed by *staccato* articulation.

We formulate the prediction task as multi-label classification since a musical passage is often associated with multiple techniques. Among the 222 recordings, we have 40 labeled with two techniques and two recordings with three techniques. Note that besides scales and arpeggios, few other techniques exist in their pure form (i.e. an entire music passage of trills). Thus we aim to identify the most prominent technique present in the recording.

5. RESULTS

5.1 Expertise Ranking

We train the projection module in 2-way and 4-way ranking as described in Section 3.1.1, and show results in Table 3 (left). For 2-way ranking, we achieve up to 93.56% accuracy, indicating a clear distinction between recordings of varying levels of expertise in most cases. Audio-MAE outperforms the other three audio encoders while Jukebox embeddings contain the least information for discerning the level of playing. The result of 4-way prediction is similar with Audio-MAE performing the best with 84% accuracy, indicating a good capability to distinguish larger expertise differences (beginner vs. virtuoso) from smaller ones. The baseline spectrogram achieves much lower metrics on both classifications, indicating that the pre-trained encoders capture more relevant nuances of musical performance. However, we are also aware that the three levels of data differs not only on performance but also on repertoire and recording environment. The effect of fine-tuning with solo piano domain data is not salient in this task: the fine-tuned Audio-MAE achieved roughly the same performance while DAC actually declined.

	Expertise Ranking				Difficulty Estimation				Technique Identification				
	2-way		4-way		9-way		3-way			Multi		Single	
	Acc	F1	Acc	F1	Acc_0	Acc_1	Acc_0	F1	mAP	AUC	Acc	Acc	F1
pre-trained													
Spec	75.90	74.73	52.34	49.94	32.98	59.17	67.21	66.75	57.49	**71.13**	73.02	**46.67**	**39.26**
Jukebox	84.51	83.79	60.41	56.75	33.41	55.36	60.49	58.27	49.33	59.79	73.33	25.44	23.53
Audio-MAE	93.48	**92.84**	**84.21**	**81.20**	31.60	**66.09**	79.03	75.11	60.69	67.51	77.46	42.22	39.81
MERT	89.48	88.73	82.12	78.81	26.55	62.28	73.13	71.38	55.76	69.05	**79.37**	37.78	35.85
DAC	86.84	87.91	77.77	76.24	27.61	59.86	69.64	69.87	48.50	57.87	78.73	24.44	23.61
fine-tuned													
Audio-MAE	**93.56**	90.22	82.26	77.82	**33.67**	60.21	77.73	**75.84**	**61.81**	67.61	79.05	35.56	33.73
DAC	82.87	81.83	78.41	76.23	28.63	61.59	64.45	62.34	50.77	59.80	79.68	26.67	25.66

Table 3. From left to right: results of 2-way and 4-way expertise ranking, 9-way and 3-way difficulty estimation, multi-label and single-label technique prediction. Best results are highlighted in bold.

5.1.1 Discussion: How far are we from predicting the Chopin Competition Winner?

From the trained 2-way expertise ranking module, we apply the *ICPC-2015* pairs as a testing set as described in Section 4.1.1. Ideally, the model should discern the three levels of piano expertise by identifying specific nuances in performance that distinguish, for example, virtuosi from advanced students. Such insights could then be applicable to evaluating competition-level performances. In Table 4, "fitting" indicates that we first fit the trained model on half of the candidates' pairs for 5 epochs and test on the other half. Without fitting, we only evaluate on these same testing pairs using the model trained in Section 5.1.

	w/o. fitting		w. fitting	
	Acc	F1	Acc	F1
pre-trained				
Spec	52.91	52.79	49.27	49.04
Jukebox	46.63	44.92	48.27	47.13
Audio-MAE	56.86	55.86	59.08	58.76
MERT	49.07	46.78	53.05	52.54
DAC	42.17	41.71	53.67	53.67
fine-tuned				
Audio-MAE	54.32	50.14	54.89	49.81
DAC	**60.49**	**60.27**	**59.87**	**59.84**

Table 4. 2-way paired-ranking test result for the competition dataset *ICPC-2015*.

Several interesting observations are made from this experiment: 1) Transferring the learnt expertise ranking into assessing competition-level playing (which shall all belongs to the virtuoso tier within our training) is challenging, considering random guess baseline of 50% accuracy in predicting the better performer within a pair. The best we achieve is slightly above 60%, possibly because that the outcomes of competitions often transcend mere audio content to include performative expression like motions, resulting in a *sight over sound* phenomenon [53]. 2) Adaptation on the competition set does not significantly boost the performance. For the pre-trained embeddings the accuracies slightly increase after fitting, but it has no effects on the fine-tuned embeddings. 3) The fine-tuned DAC embeddings, despite having a lower performance in the ranking task with three levels,

largely outperform other models in ranking the candidates in a competition setting.

Using the paired prediction results from best model (fine-tuned DAC w/o. fitting), we translate pair-wise predictions into a global ranking. Each candidate is ranked by how many wins they obtain in the 'paired matches'. Each candidate is involved in 272 pairs, given 137 candidates and we infer on each pair (136) and its inverse. Figure 2 shows the relationship between our predicted candidate win counts and the preliminary round pass hit-rate (i.e. what proportion of candidates actually passed the preliminary round). *Michał Szymanowski* is the predicted best candidate who wins in the most pairs. Overall, there exists a good correlation between our predicted win counts rank and candidates' ground truth performance: the top 18 predicted candidates all passed the preliminary round, with many of them progressing into round 2 or 3 (demonstrated by the color in Figure 2). Down to the cut-off threshold of half of the candidates, 65% of them passed the preliminary round. Finalists, however, are not necessarily predicted accurately: the winner *Seong-Jin Cho* only "wins" 39 matches and is placed towards the end in this rank, as is the third placed *Kate Liu*. Only *Charles Hamelin* (2nd place) and *Dmitry Shishkin* (6th) are placed relatively high in our ranking.
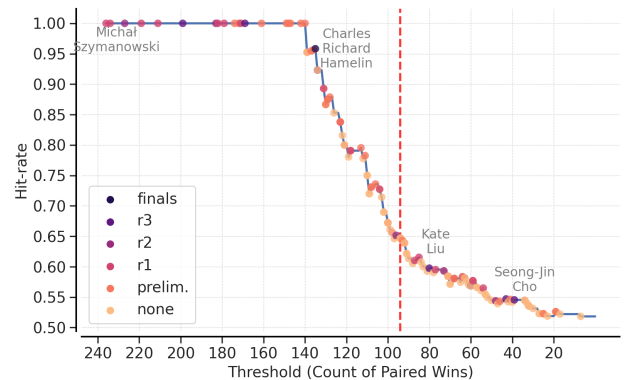

Figure 2. Paired win count threshold vs. hit-rate for preliminary round pass prediction. Each candidate is a data point colored by their ground-truth result tier (i.e. the highest round they progressed from). The red dashed line is the cut-off of half the candidates.

5.2 Difficulty estimation

The difficulty estimation experiments are defined in Section 3.1.2 on the *CIPI* dataset. In Table 3 (middle), we report the *accuracy within n* (Acc_n) for the 9 class prediction. Defined as Eq. 2, Acc_n aligns with the ordinal nature of the task and we observe results for $n = 0$ (exact match) and $n = 1$ (allowing for one-class deviation).

$$Acc_n = \frac{1}{|C|} \sum_{c \in C} \frac{|\{y \in S_c : |\hat{f}(x) - c| \leq n\}|}{|S_c|} \qquad (2)$$

The Audio-MAE embeddings yield overall the best performance for both 9-way and 3-way estimation. But it is worth noting that the untrained spectrogram baseline actually achieved accuracy metrics on-par with the audio encoders (32.98% vs. 33.67% in 9-way estimation), even higher than the worst-performing embedding of DAC.

The best we achieve with 9-class Acc_0 is 33.67% (compared to the same-set symbolic data baseline [36] of 39.47%). However, this is based on the fact that our audio embeddings are capped to 5 minutes, removing the effect of the major feature of piece length. For the 3-way classification we achieved accuracy that is on-par with the symbolic baseline, with the best Acc of 79.03%, demonstrating that the complexity of piano repertoire can also be encoded with the current pre-trained representation.

Interestingly, the Acc_0 and Acc_1 metrics do not improve hand-in-hand: Jukebox embeddings achieved the highest Acc_0 among the pre-trained models, but performed worst on Acc_1 since its prediction is sparse and scattered from observing the confusion matrix. The fine-tuned models exhibit a modest enhancement in performance metrics, as in the Acc_0 for Audio-MAE and 3-way *F1* for DAC, also helped with better generalization and less overfitting.

5.3 Technique identification

The technique identification experiment is performed as both multi-label and single-label prediction, as formulated in Section 3.1.3. In Table 3 (right), we report the mean-Averaged-Precision (mAP) and Area Under the Receiver Operating Characteristic Curve (AUC). The former accounts for the balance between precision and recall, while the latter computes area under the false positive rate and true positive rate (recall) which reflects the influence of the true negatives. We also note the multi-label accuracy Acc which accounts for all binary predictions of each class.

The most important observation on the result is that the spectrogram representation easily outperforms the audio encoder embeddings on this task, especially on the single-label prediction case (46.67%). This offers an interesting perspective on the learned embedding content: exact note onsets and texture patterns (that are associated with the piano technique classes) seem to be overlooked by the embeddings, capturing less performance-related details compared to the lower-level spectrogram. The results demonstrate that DAC and JukeBox are the least informative audio embeddings for this task (24.44% and 25.44%). Audio-MAE is the best-performing audio encoder, but the single-label prediction results do not improve with fine-tuning. On the

other hand, fine-tuning DAC on the solo piano data improved performance on this task by 2%, compared with its pre-trained version.

To gain a better understanding of the identified techniques we observe the class-wise mAP from the best-performing representation of spectrogram. As depicted in Figure 3, Repeated Notes emerge as the most accurately identified technique. Conversely, the Staccato class exhibits a decline in performance throughout the training, hinting at a potential acoustic overlap with Repeated Notes, as suggested by prior research [54]. Meanwhile, the precision for other techniques shows consistent improvement during training, achieving 40% to 60% even in more distinct technique categories like Scales and Arpeggios. However, with the highest accuracy for single-label 7-way prediction being 46.67%, it is clear that the model's ability to pinpoint techniques could be further refined, especially considering these are discernible to the human ear.
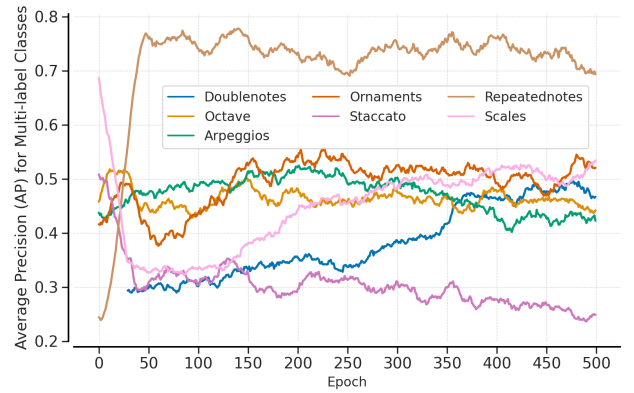

Figure 3. Average Precision for each class over epochs in multi-class prediction, from spectrogram representation.

6. CONCLUSION

Our research aimed to extend the capabilities of audio encoding models to the domain of solo piano performance understanding. Through this effort, we addressed tasks such as expertise ranking, difficulty estimation, and solo piano technique detection. The study introduced the Pianism-Labelling Dataset (PLD) and utilized a range of pre-trained audio encoders for evaluation. The curated set of performance-related attribute labeling can be contributed to multi-task learning or contrastive learning in the future.

Our results, with the highest accuracy of 93.6% in expertise ranking, suggest that models like Audio-MAE hold promise for assessing aspects of musical performance, while the codified representations such as DAC or Jukebox struggle with capturing performance nuances. However, the studies on difficulty and especially techniques suggest the limitations of current pre-trained representations in capturing pianistic textures and patterns, as they fail to outperform the spectrogram baseline, prompting for the design of a performance-oriented audio representation. Meanwhile, the case study on the Chopin Piano Competition via transferring the assessment objective confirmed that we are still far from capturing the nuances of top-level human performance.

7. ACKNOWLEDGEMENT

8. REFERENCES

[1] F. Korzeniowski and G. Widmer, "End-to-end musical key estimation using a convolutional neural network," in *25th European Signal Processing Conference, EU-SIPCO*, 2017.

[2] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," *Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 637–644, 2018.

[3] H. Flores Garcia, A. Aguilar, E. Manilow, and B. Pardo, "Leveraging hierarchical structures for few-shot musical instrument recognition," in *Proceedings of the 22th International Society for Music Information Retrieval Conference, ISMIR 2021*, 2021.

[4] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," *Proceedings of the 22th International Society for Music Information Retrieval Conference, ISMIR 2021*, 2021.

[5] M. C. McCallum, F. Korzeniowski, S. Oramas, F. Gouyon, and A. F. Ehmann, "Supervised and unsupervised learning of audio representations for music understanding," in *Proceeding of the 21t International Society on Music Information Retrieval (ISMIR)*, 2022.

[6] J. Gardner, S. Durand, D. Stoller, and R. M. Bittner, "LLark: A multimodal foundation model for music," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

[7] S. Doh, K. Choi, J. Lee, and J. Nam, "LP-MusicCaps: LLM-based pseudo music captioning," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[8] C. Palmer, "Music performance," *Annual Review of Psychology*, vol. 48, 1997.

[9] P. Ramoneda, N. Can Tamer, V. Eremenko, X. Serra, and M. Miron, "Score difficulty analysis for piano performance education based on fingering," in *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[10] P. Parmar, J. Reddy, and B. Morris, "Piano skills assessment," in *IEEE 23th International Workshop on Multimedia Signal Processing (MMSP)*, 2021.

[11] W. Wang, J. Pan, H. Yi, Z. Song, and M. Li, "Audio-based piano performance evaluation for beginners with convolutional neural network and attention mechanism," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 1119–1133, 2021.

[12] P. Seshadri and A. Lerch, "Improving music performance assessment with contrastive learning," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[13] J. Huang and A. Lerch, "Automatic assessment of sight-reading exercises," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[14] A. Morsi, K. Tatsumi, A. Maezawa, T. Fujishima, and X. Serra, "Sounds Out of Pläce? Score-independent detection of conspicuous mistakes in piano performances," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[15] H. Kim, P. Ramoneda, M. Miron, and X. Serra, "An overview of automatic piano performance assessment within the music education context," *International Conference on Computer Supported Education, CSEDU - Proceedings*, vol. 1, 2022.

[16] A. Morsi, H. Zhang, A. Maezawa, S. Dixon, and X. Serra, "Simulating and validating piano performance mistakes for music learning context," in *Sound and Music Computing Conference (SMC)*, 2024.

[17] A. Lerch, C. Arthur, A. Pati, and S. Gururani, "An interdisciplinary review of music performance analysis," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 221–245, 2020.

[18] H. Zhang, E. Karystinaios, S. Dixon, G. Widmer, and C. E. Cancino-Chacón, "Symbolic music representations for classification tasks: A systematic evaluation," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, Milan, Italy, 2023.

[19] C. E. Cancino-Chacón, "Computational modeling of expressive music performance with linear and non-linear basis function models," Ph.D. dissertation, Johannes Kepler University Linz, 2018.

[20] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao, "Multimodal foundation models: From specialists to general-purpose assistants," *arXiv preprint arXiv:2309.10020*, 2023.

[21] M. Won, Y.-n. Hung, and D. Le, "A foundation model for music informatics," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[22] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi,

W. Huang, Z. Wang, Y. Guo, and J. Fu, "MERT: Acoustic music understanding model with large-scale self-supervised training," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

[23] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved RVQGAN," in *Proceedings of the Conference on Neural Information Processing Systems (Neurips)*, 2023.

[24] P.-y. H. Hu, X. Juncheng, C. Feichtenhofer, and M. Ai, "Masked autoencoders that listen," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[25] K. Kosta, O. F. Bandtlow, and E. Chew, "Dynamics and relativity: Practical implications of dynamic markings in the score," *Journal of New Music Research*, vol. 47, no. 5, pp. 438–461, 2018.

[26] H. Kim and X. Serra, "DiffVel : Note-level midi velocity estimation for piano performance by a double conditioned diffusion model," in *16th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, Tokyo, Japan, 2023.

[27] M. Grachten, W. Goebl, S. Flossmann, and G. Widmer, "Phase-plane representation and visualization of gestural structure in expressive timing," *Journal of New Music Research*, vol. 38, no. 2, pp. 183–195, Jun. 2009.

[28] Z. Shi, "Computational analysis and modeling of expressive timing in chopin mazurkas," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[29] S. R. M. Rafee, G. Fazekas, and G. A. Wiggins, "Performer identification from symbolic representation of music using statistical models," in *Proceedings of the International Computer Music Conference (ICMC)*, 2021.

[30] Y. Zhao, C. Wang, G. Fazekas, E. Benetos, and M. Sandler, "Violinist identification based on vibrato features," in *European Signal Processing Conference*, 2021.

[31] C. Sales, P. Wang, and Y. Jiang, "An interactive tool for exploring score-aligned performances: Opportunities for enhanced music engagement," *ACM International Conference Proceeding Series*, pp. 30–33, 2023.

[32] A. Morsi, H. Zhang, A. Maezawa, S. Dixon, and X. Serra, "Simulating piano performance mistakes for music learning," in *Proceedings of the Sound and Music Computing Conference (SMC)*, 2024.

[33] H. Zhang, Y. Jiang, T. Jiang, and P. Hu, "Learn by referencing: Towards deep metric learning for singing assessment," in *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[34] M. Matsubara, R. Kagawa, T. Hirano, and I. Tsuji, "CROCUS: Dataset of musical performance critiques," in *In Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*, 2021.

[35] Y. Jiang, "Expert and novice evaluations of piano performances : Criteria for computer-aided feedback," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[36] P. Ramoneda, D. Jeong, V. Eremenko, N. C. Tamer, M. Miron, and X. Serra, "Combining piano performance dimensions for score difficulty classification," *Expert Systems with Applications*, 2024.

[37] C. Wang, V. Lostanlen, E. Benetos, and E. Chew, "Playing technique recognition by joint time frequency scattering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[38] N. Srivatsan and T. Berg-kirkpatrick, "Checklist models for improved output fluency in piano fingering prediction," in *Proceeding of the 23rd International Society on Music Information Retrieval (ISMIR)*, 2022.

[39] C. V. Hall and J. T. O'Donnell, "Calibrating a bowing checker for violin students," *Journal of Music, Technology Education*, vol. 3, no. 2-3, pp. 125–139, 2011.

[40] P. Ramoneda, M. Lee, D. Jeong, J. J. Valero-Mas, and X. Serra, "Can audio reveal music performance difficulty? insights from the piano syllabus dataset," pp. 1–13, 2024.

[41] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "Audiolm: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2023.

[42] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[43] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *Computing Research Repository (CoRR)*, 2020.

[44] E. Manilow, P. O'Reilly, P. Seetharaman, and B. Pardo, "Source separation by steering pretrained music models," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2022-May, 2022, pp. 126–130.

[45] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding LLaMA: Advancing text-to-music generation with question answering and captioning," in *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[46] J. Liang, "Acoustic prompt tuning: Empowering large language models with audition capabilities," *arXiv preprint arXiv:2312.00249*, 2023.

[47] H. F. Garcia, P. Seetharaman, R. Kumar, and B. Pardo, "VampNet: Music generation via masked acoustic token modeling," in *Proceeding of the 24th International Society on Music Information Retrieval (ISMIR)*, 2023.

[48] K. C. Puvvada, N. R. Koluguri, K. Dhawan, J. Balam, and B. Ginsburg, "Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition," in *Proceeding of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[49] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C. Z. A. Huang, S. Dieleman, E. Elsen, J. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the Maestro dataset," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019, pp. 1–12.

[50] H. Zhang, J. Tang, S. Rafee, S. Dixon, and G. Fazekas, "ATEPP: A dataset of automatically transcribed expressive piano performance," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Bengaluru, India, 2022.

[51] V. Konz, W. Bogler, and V. Arifi-M, "Saarland music data," *Late-Breaking and Demo Session of the International Society on Music Information Retrieval (ISMIR)*, 2011.

[52] G. Sandor, *On Piano Playing: Motion, Emotion and Expression*, 1981.

[53] C.-J. Tsay, "Sight over sound in the judgment of music performance," *Proceedings of the National Academy of Sciences*, 2013.

[54] R. Bresin and G. Umberto Battel, "Articulation strategies in expressive piano performance analysis of legato, staccato, and repeated notes in performances of the Andante movement of Mozart's sonata in g major (k.545)," *Journal of New Music Research*, vol. 29, no. 3, pp. 211–224, 2000.