IMPROVING MUSIC PERFORMANCE ASSESSMENT WITH CONTRASTIVE LEARNING

Pavan Seshadri
Center for Music Technology
Georgia Institute of Technology
pseshadri9@gatech.edu

Alexander Lerch
Center for Music Technology
Georgia Institute of Technology
alexander.lerch@gatech.edu

ABSTRACT

Several automatic approaches for objective music performance assessment (MPA) have been proposed in the past, however, existing systems are not yet capable of reliably predicting ratings with the same accuracy as professional judges. This study investigates contrastive learning as a potential method to improve existing MPA systems. Contrastive learning is a widely used technique in representation learning to learn a structured latent space capable of separately clustering multiple classes. It has been shown to produce state of the art results for image-based classification problems. We introduce a weighted contrastive loss suitable for regression tasks applied to a convolutional neural network and show that contrastive loss results in performance gains in regression tasks for MPA. Our results show that contrastive-based methods are able to match and exceed SoTA performance for MPA regression tasks by creating better class clusters within the latent space of the neural networks.

1. INTRODUCTION

Within the context of western classical music, musical performances are a sonic interpretation of a written musical score. Performers are tasked with interpreting the score and translating it to an acoustic rendition. In doing so, they craft a unique performance by controlling and varying performance parameters such as tempo and timing, dynamics, intonation, and tone quality [1]. These performance parameters and their variation impact the way in which listeners perceive the music, and let them distinguish between performances of the same musical score [2].

For music performers, the journey to competence and mastery often spans years of practice and tailored instruction. As music performance is an inherently complex and subjective task, proper feedback on performances is imperative to growth as a performer, and such, regular feedback by professional musicians is necessary. Music teachers are expected to evaluate students on various criteria such as musicality or tone quality, although rating these criteria is highly subjective, complicating the task of consistent and objective Music Performance Assessment (MPA) [3,4]. These challenges, however, do not reduce the need of assessing music performances, e.g., in institutions of musical education. Thus, any effort towards either formalizing human assessments or the creation of objective, reproducible, and unbiased systems for automatic assessment contributes to overcoming the above-mentioned challenges.

A system for automatic MPA can be used for software-based music tutoring applications to allow for easier accessibility of music education and individualized instruction. Past this, such objective assessment systems might also serve as tools for the evaluation of performance generation systems.

The approaches in automatic MPA follow the same general historical patterns as other audio analysis systems. Older systems extract hand-crafted features from recorded performances and then use a data-driven approach such as a regression model to map the features to a grade or assessment rating that reflects human ratings [5]. Deep learning methods have since been found to outperform feature extraction-based methods [6]; however, modern systems still fall short of the reliability required for a ready-to-use system [7].

Representation learning aims to accurately encode relevant and useful characteristics into a compressed representation. Representation learning methods such as VGGish have been shown to encode powerful audio features into a compressed representation which —when used as input to classification systems— can produce state of the art performance [8]. Contrastive learning is an emerging representation learning method which uses a distance-based loss between pairs of encoded training points in order to create meaningful class separation within the latent space of a neural network [9].

This study aims to investigate the use of contrastive learning to improve the performance of MPA systems. We investigate the use of contrastive-based learning methods in a regression task, where a deep neural network taking an input audio recording of a musical performance is tasked with estimating a numerical rating consistent with that of a professional judge. Our hypothesis is that learning a structured latent space will improve the ability of the regression components of MPA models in predicting an assessment rating. We investigate two methods of incorporating con-

Attribution: P. Seshadri and A. Lerch, “Improving Music Performance Assessment with Contrastive Learning”, in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.,* Online, 2021.

trastive learning into a standard CNN-based architecture to learn a structured latent space, (i) a two-step training method introduced by Khosla et al. [10], and (ii) a joint loss method combining the contrastive loss term with a mean squared error loss term, a standard loss for training regression systems. As contrastive loss within a supervised context is generally designed for classification tasks [10] as opposed to regression tasks, we introduce a weighted contrastive loss suitable for regression tasks.

The remainder of this paper is structured as follows. First, we give an overview of music performance assessment and previous work on contrastive loss. Then, we introduce the proposed method in Sect. 3. Section 4 introduces our experimental setup. In the following Sect. 5, we evaluate the performance of the contrastive-based methods against a baseline architecture to predict a regression rating and perform analysis on the latent space of the baseline and contrastive-based methods to determine the efficacy of this clustering on the overall performance. Overall, we find that contrastive-based learning is able to better cluster the latent representation and produce performance gains within our MPA regression task.

2. RELATED WORK

MPA aims to understand and model the parameters of a musical performance and investigate their impact on a human listener [11]. MPA systems thus have the goal of assessing musical performances based on audio recordings without the input of expert judges. Early research on musical performances centered around analyzing symbolic data extracted from MIDI devices [12, 13], whereas recent research has increasingly focused on analyzing raw audio [11, 14]. In human performance assessment, music instructors must discern the individual subjective qualities and criteria and their importance. Similarly, automatic performance assessment systems extract features representing the audio file and then use a data-driven model to estimate the assessment rating. The features are either hand-crafted for the task [5, 15–18], or learned from the training data [6, 19, 20]. Systems with handcrafted features often use traditional machine learning approaches [21] while feature learning is usually done within a more complex neural model with low-level input representations such as spectrograms [22].

Some studies specifically aim to automatically produce a numerical rating on a predefined scale from audio representations of a musical performance, which involves implicitly learning the aspects of performances that correlate to certain rating criteria [6, 7, 20]. However, since numerical scores do not inherently include specific performance feedback, understanding the impacting factors can be challenging. The methods based on deep neural networks, while generally yielding superior performance, usually lack interpretability. Learning a structured latent space is a first step towards having a more easily understandable representation. Representation learning is an emerging method for performance assessment. For example, Huang et al. proposed a joint-embedding network which learns a shared latent space of a performance and its written score and derives a regression rating by the cosine similarity between the two embeddings [7]. Representation learning methods thus potentially provide both performance improvements, as well as better interpretability of numerical scoring models, such as the ones in this study.

An emerging method of representation learning is the Contrastive Loss. Contrastive Loss aims to regularize the latent space so that the distances between latent vectors are meaningful. This is achieved by comparing the distances between the latent representations of pairs of training points and pushing them within a set distance in the latent space if they have similar labels, and outside this set distance if they are dissimilar. These distances are compared within the contrastive loss function of a model in order to encode the information within the latent vectors. This ideally creates class clusters within the latent space. Contrastive-based loss functions are often used to specifically learn structured latent representations of data [9, 10, 23], which then can be adapted for downstream tasks by training classifiers on these produced latent vectors [10, 23]. There has been considerable work done on the use of a supervised contrastive loss to cluster latent spaces for classification tasks. Chopra et al. introduce the max margin contrastive loss, and discuss its potential to discriminate classes when the exact number of classes may not be known, such as within recognition or verification tasks [9]. Khosla et al. investigate a supervised contrastive loss to train deep neural networks for classification tasks on the ImageNet dataset, and found that it outperforms general cross entropy based methods [10]. This implies that using the contrastive loss can produce an advantageous latent space layout more suitable for the following tasks. Ferraro et al. investigate the use of contrastive learning for music and audio for three downstream MIR tasks, genre classification, playlist continuation, and automatic tagging and found that contrastive-based learning outperforms the baseline within all three tasks and achieves comparable performance to SoTA [23]. Their findings suggest that contrastive learning is able to cluster similar musical recordings within the latent space of deep neural networks. To our knowledge, the use of contrastive learning has not been investigated within the context of MPA. Since it has been found to be advantageous within classification tasks across several modalities [9, 10, 23], we study the application of contrastive learning to MPA.

3. METHOD

We propose a weighted contrastive loss as a modification of the max margin contrastive loss introduced by Chopra et al. [9]. The loss function is adapted to be suitable for regression tasks. We investigate incorporating the contrastive loss via two different training scenarios for a convolutional neural network architecture. [1]

[1] The code is available at: https://github.com/pseshadri9/contrastive-music-performance-assessment, last accessed 8/3/2021.

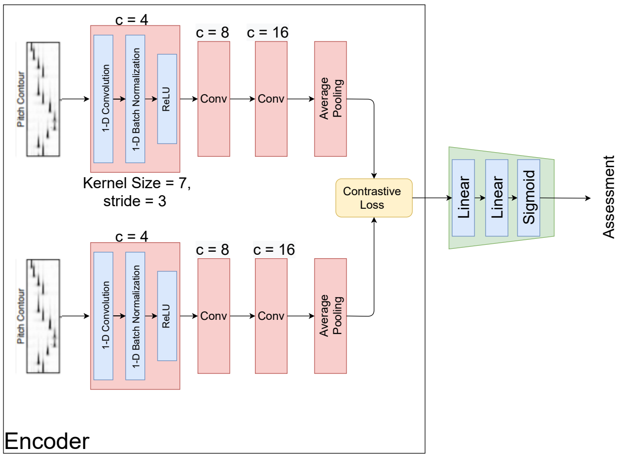


Figure 1: Contrastive-Based network architecture for regression.

3.1 Network architectures

3.1.1 *Baseline*

The baseline network used is the PCConvNet architecture introduced by Pati et al. [6]. This architecture takes pitch contours as input and uses three convolutional layers followed by an average pooling layer in order to predict assessment ratings. Each convolutional layer contains a 1-D convolution, 1-D batch normalization [24], and ReLU nonlinearity.

3.1.2 *ContrastiveCNN*

Based on the baseline system, we design the network visualized in Figure 1. Each branch of the network uses the same convolutional layers of the baseline with shared weights. Two linear layers are appended to this to predict the final rating with a sigmoid activation.

A two-step training procedure as detailed by Khosla et al. [10] is used to train this model. First, the encoder is trained using contrastive loss over the output latent vectors. After this, the encoder weights are frozen, and the linear layers are trained to regress this space using a mean squared error loss. Each datapoint within a training pair is fed through one encoder channel.

3.1.3 *ContrastiveCNN-JL*

The architecture of this network is equivalent to the ContrastiveCNN, but differs in training procedure. Rather than the two-step training procedure outlined above, the loss L is the addition of the contrastive loss L_{C} over the latent vectors and the mean squared error loss L_{MSE}

$$L = L_{\text{MSE}} + L_{\text{C}}. \tag{1}$$

3.1.4 *Input representation*

The input for each model is a pitch contour of each individual audition. Each pitch contour is an $N \times 1$ vector representing the fundamental frequency of each chunk of a performance of sequence length N chunks. Pitch Contour representations were extracted using the pYIN algorithm [25] at a sample rate of 44.1 kHz with a block size and hop size of 1024 and 256 samples, respectively. The extracted fundamental frequencies are converted to MIDI pitch values and normalized to a range of [0,1] by dividing by 127, the maximum MIDI value.

	Middle School	Symphonic Band
Alto Sax	696	641
Clarinet	925	1156
Flute	989	1196

Table 1: Number of recordings per instrument.

3.2 Weighted Contrastive Loss

Contrastive loss is generally used for classification tasks in order to create distinct class boundaries within the latent space [9, 10, 23, 26]. The standard max margin contrastive loss [9] is defined as:

$$L_{\text{C}} = \frac{1}{2} Y D^2 + \frac{1}{2}(1 - Y)\max(0, (m - D))^2, \tag{2}$$

where $Y = 1$ if the two datapoints in the pair have the same ground truth label, and $Y = 0$ if they do not. D is the Euclidean distance between the two latent vectors, and m is a set margin distance for which similarly labeled points should be clustered within. This results in points from the same class clustered together, while differently labeled points will be pushed past this pre-defined distance margin.

Since this loss is not suitable for regression tasks like ours, we propose a weighted contrastive loss term. For this new loss, we first split our continuous regression range [0, 1] into C evenly spaced rating bins. For $C = 5$, for example, each rating bin has a range of 0.2, with exact multiples of 0.2 serving as the lower bounds for each bin X (i.e., $[0, 0.2)$, $[0.2, 0.4)$,$\ldots$). Each datapoint is assigned its respective bin according to its ground truth rating. These bins are then assigned the class indices $[0, 1, 2, \ldots, C-1]$, which will be used for our weighted contrastive loss. Second, we propose a variable margin to represent the ordered nature of the rating bins. For example, it is expected that the rating bin spanning $[0, 0.2)$ should have a greater distance from the bin covering $[0.8, 1]$ than from the $[0.2, 0.4)$ bin, as the bins themselves express a rating distance. The variable margin can therefore be defined as

$$m = |X_i - X_j| \cdot s, \tag{3}$$

where X_i and X_j represent the ground truth class indices of each datapoint within a pair (X_i, X_j) and s is the set margin distance. This variable margin scales the set distance proportionally to the expected distance between each rating bin. This variable margin then replaces the fixed margin m in Eq. (2).

4. EXPERIMENTS

Our experiments investigate primarily the performance of the proposed contrastive-based methods for MPA. In particular, we are interested in evaluating (i) the raw performance

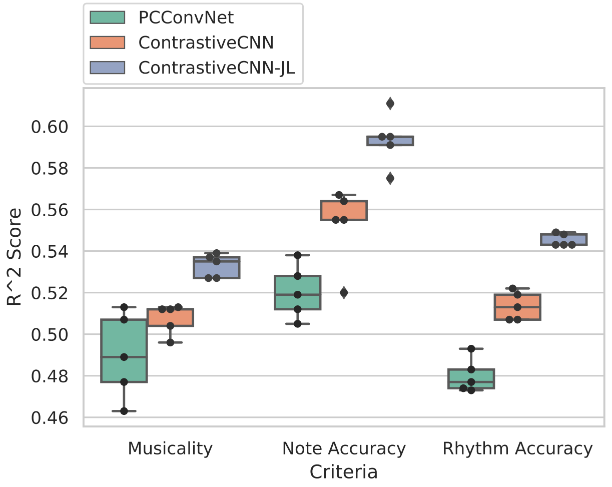


Figure 2: Results over the *Middle School* set.

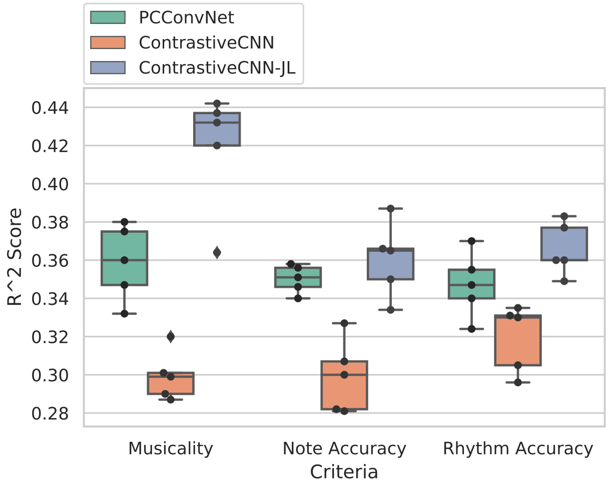


Figure 3: Regression results over the *Symphonic Band* set.

in predicting ratings, (ii) the quality of the clustering of the latent spaces, and (iii) the effect of this clustering on the performance. We evaluate the learned representation both quantitatively and qualitatively.

4.1 Dataset

The used dataset comprises audio recordings and ratings of auditions from the Florida Bandmaster's Association (FBA) from 2013 to 2018. This dataset contains raw audio recordings from three different levels of all-state auditions, *Middle School*, *Concert Band*, and *Symphonic Band*. Each student performs a prepared *lyrical exercise*, *technical exercise*, *scales*, and a *sight reading exercise*. This dataset includes several monophonic and percussive instruments. A subset of these data was used in this study, using the technical exercise from the alto saxophone, Bb clarinet, and flute recordings for the *Middle School* and *Symphonic Band* levels. Table 1 shows the number of recordings per instrument for both *Middle School* and *Symphonic Band*. The average duration of a *Middle School* and a *Symphonic Band* recording is approximately 30 s and 50 s, respectively.

Each singular recording represents the complete audition for one student and has assessment ratings by an expert judge for four assessment criteria defined by the FBA: *musicality*, *note accuracy*, *rhythm accuracy*, and *tone quality*. For consistency, we normalized each rating to the range [0, 1] by dividing the maximum rating, with 0 representing the worst possible score, and 1 representing the best possible score. Furthermore, the tone quality rating was ignored for this study as the audition is represented as pitch contours at our network input, a representation that does not carry sufficient information for modeling this criterion.

4.1.1 Pre-processing

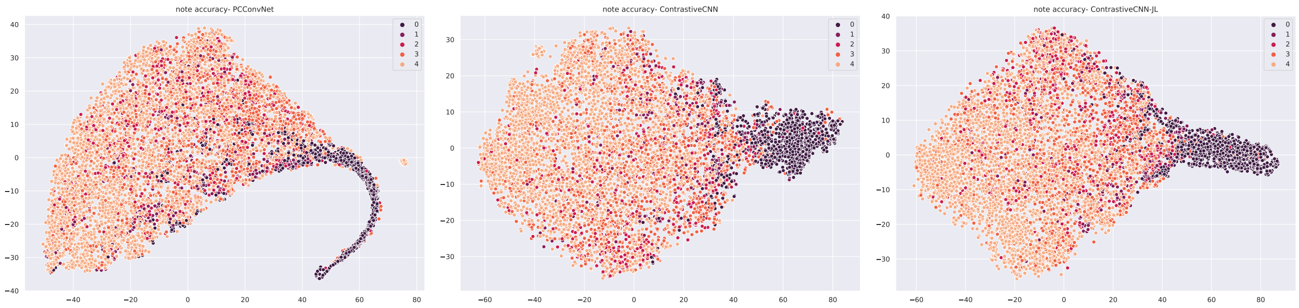
Pitch contour representations were computed from raw audio recordings. Data augmentation via random chunking was used while training due to its ability to improve model performance [6]. Each pitch contour is chunked into sections of length 1000 (about 6 s) by randomly selecting the start position. This approach has been shown to improve model performance [6]. We assume the chunked segment would receive the same assessment rating as the entire audition recording.

4.2 Training procedures

Pairs for the contrastive loss were randomly sampled via generated random sequences each batch. Each datapoint within the pair was fed into a separate encoder channel of the model. Each model was trained using a stochastic gradient descent optimizer with a weight decay of 1e-5 and momentum of 0.9. Early stopping was applied in each training sequence to stop if the validation loss had not decreased in 75 epochs. For training and evaluation, each dataset was split into training, testing, and validation sets in an 8:1:1 ratio. To measure the variance of the models, each model was trained five times using random seeds, as represented by the box plots. Within the two step training method following [10], the encoder channels were trained for 150 epochs at a learning rate of 0.1, while the linear layers were trained for 300 epochs at a learning rate of 0.005. The Joint Loss Network was trained for 300 epochs at a learning rate of 0.005.

4.3 Evaluation

We investigate the performance of the baseline and the contrastive-regularized networks amongst three different rating criteria, *musicality*, *note accuracy*, and *rhythm accuracy*. We predict the ratings for these criteria over both the *Middle School* and *Symphonic Band* dataset to determine performance over different levels of musical complexity, which can provide insights over the performance of MPA systems as musical complexity increases. The *Concert Band* dataset was omitted for consistency, as it was not evaluated in previous MPA studies that used this dataset [6, 7]. Each model (PCConvNet, ContrastiveCNN, ContrastiveCNN-JL) was trained separately for each assessment criterion on both datasets. The unaltered PCConvNet [6] served as the baseline model.



(a) PCConvNet

(b) ContrastiveCNN

(c) ContrastiveCNN-JL

Figure 4: T-SNE visualization of the latent space of the three presented models.

4.3.1 *Regression analysis*

The coefficient of determination $(R^2 score)$ over the output scores serves as the evaluation metric:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{4}$$

where y_i is the ground truth rating for a given datapoint, $\hat{y}_i$ is the predicted rating, and $\bar{y}$ is the mean ground truth rating over the entire set.

4.3.2 *Latent space analysis*

Using each trained model, the latent vectors of the testing set were obtained by only passing the input through each model's encoder channels. A visualization was produced by applying T-SNE dimensionality reduction [27] to the latent vectors and plotting the output. Optimal parameters were found via a parameter search.

For a quantitative evaluation of the clustering quality, the latent space is evaluated by its Davies-Bouldin index [28]. The Davies-Bouldin index describes the average similarity of each cluster to its most similar cluster, which is defined as the ratio of within-cluster distances to between-cluster distances. The minimum index is zero and smaller values indicate better clustering [28].

5. RESULTS AND DISCUSSION

5.1 Regression results

Figure 2 and Figure 3 detail the results of the models on the *Middle School* and *Symphonic Band* datasets, respectively. Each box plot contains the five runs with random seeds 0-4 per each criteria and model. We can make the following observations:

(i) All models perform better on the *Middle School* set than the *Symphonic Band* set (higher R^2 score). One possible explanation for this is that the *Symphonic Band* auditions tend to be longer and more complex with higher skilled players, potentially increasing the difficulty of extracting meaningful features representing the quality of the performance.

(ii) All contrastive-based models outperform the baseline on the *Middle School* set; however, only the ContrastiveCNN-JL meets and outperforms the baseline on *Symphonic Band*. This implies that the contrastive learning is more beneficial at a lower complexity of performance, but possibly has difficulty with data of higher complexity. One possible explanation could be that with a higher level of performance, and thus a higher complexity of information within each latent vector, the contrastive loss is unable to properly semantically relate the distances to the quality of the performance.

(iii) The ContrastiveCNN-JL outperforms both the baseline and the ContrastiveCNN in every trial. This implies that the information gained by combining the traditional loss term with the contrastive loss helps learning a more meaningful latent space representation.

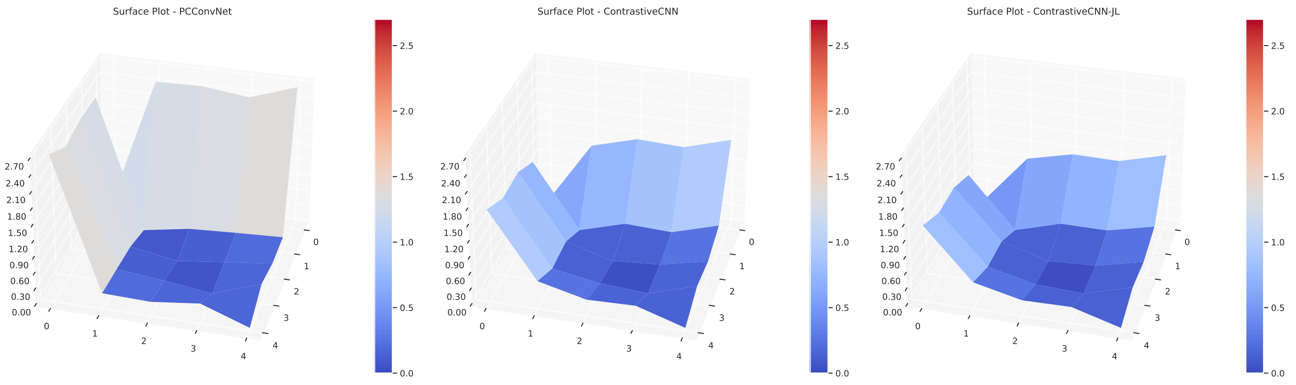
5.2 Latent space analysis

5.2.1 *T-SNE plots*

T-SNE visualizations of the latent space are presented for the baseline PCConvNet, ContrastiveCNN, and ContrastiveCNN-JL in Figure 4. As a example, we only present results for *Note Accuracy* on the *Middle School* dataset. The effect of the contrastive loss can be easily noticed, although the embedding spaces are not ordered perfectly in either case. The two models based on contrastive loss display a more defined distinction between low classes (0, 1) and higher classes (3, 4). Small same-class clusters can also be identified.

5.2.2 *Class Distance Surface plots*

Figure 5 shows the distances between the centroids of each class cluster within the latent space of the models trained on the *Middle School* dataset for *Note Accuracy*. While the contrastive-based models appear to have trouble properly ordering the middle range of ratings between classes 2 and 3, the distances appear to scale more smoothly than the distances within the baseline PCConvNet, indicating better ordering within the latent space.



(a) PCConvNet (b) ContrastiveCNN (c) ContrastiveCNN-JL

Figure 5: Class distances in the learned latent space of the three models.

5.2.3 Davies-Bouldin Index

Figure 6 shows the Davies-Bouldin indices of each model on the *Middle School* regression set. Each contrastive-based model has a considerably lower index than the baseline, which indicates that the latent space clustering is improved. Within this set, the lower the Davies-Bouldin index, the better the regression performance, implying that better clustered latent spaces do correlate with better regression.

6. CONCLUSION

This paper presented an approach to representation learning to improve the accuracy of a system for music performance assessment. We introduced a weighted contrastive loss suitable for regression tasks and showed how this latent space regularization improves results on a large real-world dataset for music performance assessment.

In future work, we plan to incorporate score information into the models, as this has been shown to improve performance [7]. More analysis should be done within contrastive learning methods to assess the effect of margin size, and number of classes on the performance of the model and the goodness of its clustering. Another approach to ensure that the learned representations contain relevant information is multi-task learning. It is worth investigating what related tasks might help increase the performance of music performance assessment. Moreover, supervised latent space regularization methods such as AR-VAE [29] and I-VAE [30] might be incorporated to force specific dimensions to specific performance characteristics.

7. ACKNOWLEDGMENTS

We would like to thank the Florida Bandmasters Association for providing the dataset used in this study.

We also gratefully acknowledge NVIDIA Corporation (Santa Clara, CA, United States) who supported this research via the NVIDIA GPU Grant program.

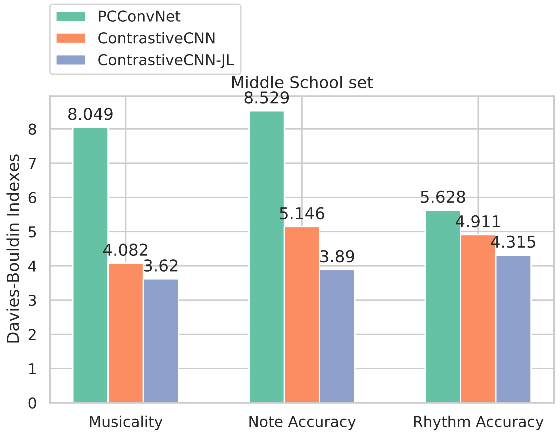


Figure 6: Davies-Bouldin indices of each model over the *Middle School* dataset. Smaller values indicate better clustering.

8. REFERENCES

[1] E. F. Clarke, "Understanding the Psychology of Performance," in *Musical Performance – A Guide to Understanding*, J. Rink, Ed. Cambridge: Cambridge University Press, 2002.

[2] A. Lerch, C. Arthur, K. A. Pati, and S. Gururani, "An Interdisciplinary Review of Music Performance Analysis," *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 3, no. 1, pp. 221–245, 2020.

[3] B. C. Wesolowski, S. A. Wind, and G. Engelhard, "Examining Rater Precision in Music Performance Assessment: An Analysis of Rating Scale Structure Using the Multifaceted Rasch Partial Credit Model," *Music Perception: An Interdisciplinary Journal*, vol. 33, no. 5, pp. 662–678, Jun. 2016.

[4] S. Thompson and A. Williamon, "Evaluating Evaluation: Musical Performance Assessment as a Research Tool," *Music Perception: An Interdisciplinary Journal*, vol. 21, no. 1, pp. 21–41, Sep. 2003.

[5] C.-W. Wu, S. Gururani, C. Laguna, A. Pati, A. Vidwans, and A. Lerch, "Towards the Objective Assessment of Music Performances," in *Proceedings of the International Conference on Music Perception and Cognition (ICMPC)*, San Francisco, 2016, pp. 99–103.

[6] K. A. Pati, S. Gururani, and A. Lerch, "Assessment of Student Music Performances Using Deep Neural Networks," *Applied Sciences*, vol. 8, no. 4, p. 507, Mar. 2018.

[7] J. Huang, Y.-N. Hung, K. A. Pati, S. Gururani, and A. Lerch, "Score-informed Networks for Music Performance Assessment," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*. Montreal: International Society for Music Information Retrieval (ISMIR), 2020.

[8] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "CNN Architectures for Large-Scale Audio Classification," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 131–135, iSSN: 2379-190X.

[9] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality Reduction by Learning an Invariant Mapping," in *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, Jun. 2006, pp. 1735–1742, iSSN: 1063-6919.

[10] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 661–18 673, 2020.

[11] A. Lerch, *Software-Based Extraction of Objective Parameters from Music Performances*. München: GRIN Verlag, 2009.

[12] C. Palmer, "Mapping Musical Thought to Musical Performance," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 15, no. 2, pp. 331–346, 1989.

[13] B. H. Repp, "Patterns of note onset asynchronies in expressive piano performance," *Journal of the Acoustical Society of America (JASA)*, vol. 100, no. 6, pp. 3917–3932, 1996.

[14] S. Dixon and W. Goebl, "Pinpointing the Beat: Tapping to Expressive Performances," in *Proceedings of the 7th International Conference on Music Perception and Cognition (ICMPC)*, Sydney, 2002.

[15] T. Nakano, M. Goto, and Y. Hiraga, "An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features," in *Proceedings of the International Conference on Spoken Langaunge Processing (INTERSPEECH)*, vol. 12, Pittsburgh, PA, 2006, p. 1.

[16] T. Knight, F. Upham, and I. Fujinaga, "The Potential for Automatic Assessment of Trumpet Tone Quality," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Miami, FL, 2011, pp. 573–578.

[17] C. Dittmar, E. Cano, J. Abeßer, and S. Grollmisch, "Music Information Retrieval Meets Music Education," *Multimodal Music Processing*, vol. 3, pp. 95–120, 2012.

[18] S. Gururani, K. A. Pati, C.-W. Wu, and A. Lerch, "Analysis of Objective Descriptors for Music Performance Assessment," in *Proceedings of the International Conference on Music Perception and Cognition (ICMPC)*, Toronto, Ontario, Canada, 2018.

[19] Y. Han and K. Lee, "Hierarchical Approach to Detect Common Mistakes of Beginner Flute Players," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Taipei, Taiwan, 2014, pp. 77–82.

[20] C.-W. Wu and A. Lerch, "Learned Features for the Assessment of Percussive Music Performances," in *Proceedings of the International Conference on Semantic Computing (ICSC)*. Laguna Hills: IEEE, 2018.

[21] A. Vidwans, S. Gururani, C.-W. Wu, V. Subramanian, R. V. Swaminathan, and A. Lerch, "Objective descriptors for the assessment of student music performances," in *Proceedings of the AES Conference on Semantic Audio*. Erlangen: Audio Engineering Society (AES), 2017.

[22] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5070–5074.

[23] A. Ferraro, X. Favory, K. Drossos, Y. Kim, and D. Bogdanov, "Enriched Music Representations With Multiple Cross-Modal Contrastive Learning," *IEEE Signal Processing Letters*, vol. 28, pp. 733–737, 2021.

[24] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the International Conference on Machine Learning*. PMLR, Jun. 2015, pp. 448–456, iSSN: 1938-7228.

[25] M. Mauch and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, May 2014, pp. 659–663, iSSN: 2379-190X.

[26] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A Survey on Contrastive Self-Supervised Learning," *Technologies*, vol. 9, no. 1, p. 2, Mar. 2021, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[27] L. van der Maaten and G. E. Hinton, “Visualizing High-Dimensional Data Using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[28] D. L. Davies and D. W. Bouldin, “A Cluster Separation Measure,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[29] K. A. Pati and A. Lerch, “Attribute-based Regularization for Latent Spaces of Variational Auto-Encoders,” *Neural Computing and Applications*, 2020.

[30] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen, “Variational Autoencoders and Nonlinear ICA: A Unifying Framework,” in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, S. Chiappa and R. Calandra, Eds., vol. 108. PMLR, Aug. 2020, pp. 2207–2217.