A Supervised Learning AI Model for Automated Holistic Vocal Performance Feedback

Saanvi Bhargava
The Harker School, 500 Saratoga Ave, San Jose, CA 95128
saanvi.bhargava@gmail.com

1. Abstract

Scalable music education requires giving fast feedback on student audio performances. Current manual feedback mechanisms are given by teachers, rendering them subjective and, therefore, sometimes inaccurate. Current technological feedback mechanisms evaluate whether a student is correct on a single note rather than the entire music piece, not providing cumulative or numerical feedback. An AI model is presented for automatically grading vocal music recordings cumulatively on pitch and rhythm given a reference piece of music. The model predicts a numerical grade for the performance of a reference piece of music. The ML model is then tested for accuracy on a dataset of corresponding audio recordings (performance and reference) and tagged human scores for these performances. Besides demonstrating the feasibility of developing an objective music grading system, the investigation presented in this paper also reveals some important limitations and subjectivity of current music grading systems, opening opportunities for future work in the community.

Keywords

Computational musicology, Music education technology, Artificial Intelligence, Equity in music education, AI music, Automated music transcription

2. Introduction

Since the 1980s, arts education has suffered, and access to high-quality music education has diminished in underprivileged areas (Hash 2021). Timely and accurate feedback is crucial to student learning and skill development. Due to a lack of teachers and resources, students are not able to receive clear and elaborate feedback, which may result in lower levels of music fluency. When students receive immediate feedback, they can identify their mistakes, understand their strengths and weaknesses, and make targeted improvements, ultimately leading to better overall education outcomes.

The problem at hand is the need for real-time detection and numerical evaluation of music errors and holistic feedback, explicitly focusing on pitch matching and rhythm accuracy.

To address this, there is a need for an automated system that can detect music errors in real-time, analyze pitch matching, and evaluate rhythm accuracy. This paper aims to build upon previous works and present a machine-learning model that provides cumulative feedback on pitch and rhythm accuracy for vocal music performances. The algorithm considers the entire music piece, analyzing the relationship between notes and evaluating the overall performance. Furthermore, my approach focuses on real-time detection and evaluation, allowing for immediate feedback to students during their practice sessions and aiming to improve grading ease for teachers.

3. Related Work

Multiple avenues have been explored in the search for an effective approach to enhance music education. The related work can be categorized into three categories –

1. Music Education Technology
2. Automatic Music Transcription
3. Rubrics for scoring music performances

3.1. Innovative Tech in Music Education

Current solutions and advances in music technology have greatly impacted music education, providing innovative tools and platforms to enhance learning experiences.

One study titled "Music Software in the Technology Integrated Music Education" (Nart, Sevan (2016)) aimed to identify beneficial software used in music education. The research explored the usage of software and its impact on music education, highlighting the importance of incorporating technology in the teaching and learning process.
Another work titled “Examination of STEAM-based Digital Learning Applications in Music Education’, European Journal of STEM Education” delved into the realm of digital applications specifically designed for music education (Özer and Demirbatır 2023). The study provided an in-depth review of several applications and categorized them based on their relevance and effectiveness in enhancing music learning.
Further, Yiting and Sonquan, in ‘Modern technology-enabled approaches in preschool music education’, aimed to address average and low knowledge levels in conventional music preschool education (Yiting and Sonquan 2022). In this paper, researchers developed a modified training program for preschoolers utilizing modern technologies. The program positively impacted the development of communication and logical skills in preschoolers, emphasizing the potential of technology in early music education.

In the realm of mobile technology-supported music education (MTSME), a systematic review examined studies from 2008 to 2019 (Liu et al 2023). The analysis revealed an increase in MTSME studies during that period, focusing on learner perceptions and the use of tablet computers. The review identified common learning strategies and provided valuable insights into the role of mobile technology in facilitating music education and its impact on learning motivation.

Other initiatives are also being created. Project Music X, an online music education initiative, utilizes web 2.0 technologies to provide accessible music programs in remote schools (Crawford 2013). By leveraging technology, students in rural areas can access high-quality music education experiences through online resources, workshops, and live concert streaming, transforming the traditional music classroom.

3.2. Automatic Music Transcription (AMT)

Automatic Music Transcription (AMT) is a popular problem being tackled in the field of computation musicology. AMT is an approach that could be considered for automatic vocal grading. Various techniques have been employed, including statistical, perceptually motivated, and unsupervised learning methods for digital signal processing (DSP) (Klapuri 2006). These

methods aim to overcome the challenges associated with AMT by leveraging different techniques and algorithms.

One thesis discusses the use of music transcription in the engineering field to simplify the creation of music scores (Gao). The project includes multiple processing steps, such as pitch detection, onset detection, and generating the transcription. The focus is on monophonic music recordings, serving as an initial attempt in the broader field of music transcription.

Benetos and Dixon present a method for automatically transcribing polyphonic music from audio recordings (Benetos et al 2019). It uses advanced techniques to estimate multiple pitches and reduce noise in the audio. The system selects the best pitch candidates and computes their harmonic properties. It also considers the overall characteristics of the music to improve accuracy. Transcribing polyphonic music could be useful when evaluating vocal music sung in different parts.

By exploring these different avenues and employing various techniques such as DSP4, machine learning, deep learning, and NMF, researchers strive to advance the field of AMT and develop more accurate and reliable transcription systems. Automatic music transcription is a greatly researched field and can prove useful for vocal performance scoring depending on the data provided.

3.3. Rubrics for Scoring Music Performances

When it comes to automatically grading vocal music, previous research has explored the development of rubrics to assess music accuracy. One study investigated the effectiveness of rubrics in evaluating vocal performances (Gynnild 2019). The researchers collected a group of teachers who discussed and created a standard rubric but concluded that a standard rubric may not be the most suitable approach.

However, they found that for entry-level students, a rubric can still be a valuable tool for assessing music accuracy.

Another relevant paper examined the implications of a graded system of assessment in the context of vocational education and training (Skiba 2020). This study reviewed a previous grading system and explored its applicability to assessing vocal music; it determined that rubrics must be specified for different purposes and for different graders. These research papers shed light on the potential benefits and limitations of rubrics for grading music accuracy, emphasizing the importance of considering the specific needs and skill levels of students when designing assessment frameworks in this domain.

Overall, these research papers highlight the diverse range of current solutions and advances in music technology, providing educators with tools, applications, and platforms to enrich music education and empower students with enhanced learning experiences. These papers also lay a great foundation for the automatic grading of vocal performances but still leave this problem unsolved.

4. Evaluation Dataset

The initial step involved identifying and accessing the MAST Melody Dataset, which proved to be a valuable resource (Bozkurt et al 2017, 2023). This dataset contained both reference (ref) piano and performance (per) singing audio files. The dataset also provided evaluation by experts on a numerical grading scale (1-4, where 4 is perfect and 1 is entirely incorrect) based on how close each "per" file was to the corresponding "ref" file. Each performance was scored by 5 judges. Some of the performance scores have consensus, and others do not. The majority score was used if there was no consensus in the scoring across judges.

The dataset was later supplemented with files of F0 scores for all ref and per audio files. F0 scores are lists of pitches sampled at certain (small) intervals. Table 1 (Total Sample Analysis) shows the breakdown of samples and their scores. The data shows that most of the samples were scored either 1 or 4.

Table 1: Total Sample Analysis.

Number of Reference Audio Files	2829
Number of Performance Audio Files	1046
Number of Audio Samples Scored as a 1	409
Number of Audio Samples Scored as a 2	167
Number of Audio Samples Scored as a 3	83
Number of Audio Samples Scored as a 4	280
Number of Audio Samples with no score	107

Table 2 shows the breakdown of the number of samples with consensus for each score category. This table indicates that most of the samples with consensus scores were either scored 1 or 4 by the judges, supporting the hypothesis that is difficult for humans to judge on a gradient.

Table 2: Sample Analysis with Score Consensus.

Number of Samples w/ consensus and score 1	196
Number of Samples w/ consensus and score 2	4
Number of Samples w/ consensus and score 3	5
Number of Samples w/ consensus and score 4	193
Number of Samples w/ consensus	398

5. Time Series Analysis of F0 Data

The F0 Data ('Fundamental frequency' (2023)) -- files of Hz values (spaced out evenly for the recordings) for each reference and performance file contained in the melody dataset are essentially time series data of pitches.

For each reference and performance file, I filtered the F0 data to remove all zeros to discard pauses or white noise. To properly compare reference and performance F0 files, it was necessary

to scale corresponding files. The reason for scaling was to account for when the singer had changed the octave or key of the piece. To scale the performance F0 data, I calculated the averages of F0 data for both files (reference and performance) and used the ratio of averages to change all pitches of one of the files to, as closely as possible, match the key of the other.

The next step was to find the similarity (distance) between the reference and performance using the filtered F0 data, which is time series data.

Many algorithms have been researched in the past to compare the similarity of time series data. I considered Euclidean Distance, MAPE, and Dynamic Time Warping algorithm ('Dynamic time warping' (2023)). The reference and performance samples can be of different lengths (different number of samples); hence, Euclidean and MAPE cannot be used for distance calculation.

DTW represents an efficient method for systematically exploring various feasible time alignments between the two time series, ultimately selecting the most suitable match. DTW can detect time-shifted and distorted variations of analogous series effectively. The DTW distance was calculated using the DTAIDistance library (Meert et al, 2020).

Running this on all files resulted in a list of numbers representing the distances between each reference and the corresponding performance. This distance between performance and reference can now be used to train a model to predict an overall score for the vocal performance. This is detailed in the upcoming scoring section of the paper.

5.1. Inconsistency in Human Judging

Before training a model, I did more statistical and visual analysis of the DTW distance for various sample sets. First, I split the data into sets by the human score (score 1, score 2, score 3 and score 4). The visual inspection (shown Figures 1-4) suggested that the DTW distance in each sample set has a high variance and many outliers. Mean and MAD (Median absolute Deviation) were calculated for all the sets, depicted in Table 3. MAD is a well-established measure for detecting outliers.

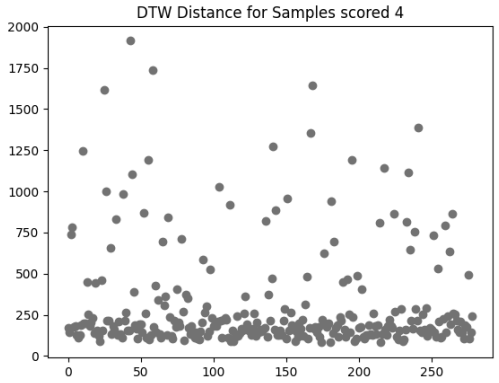

Figure 1: DTW Distance for Samples scored 4.

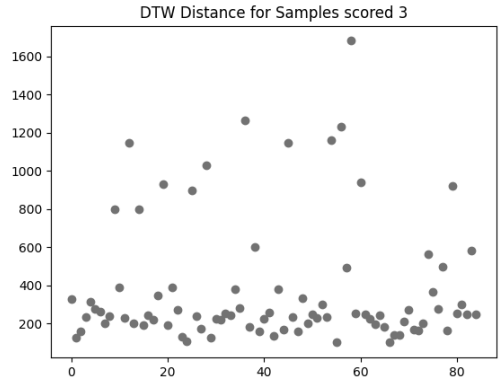

Figure 2: DTW Distance for Samples scored 3.

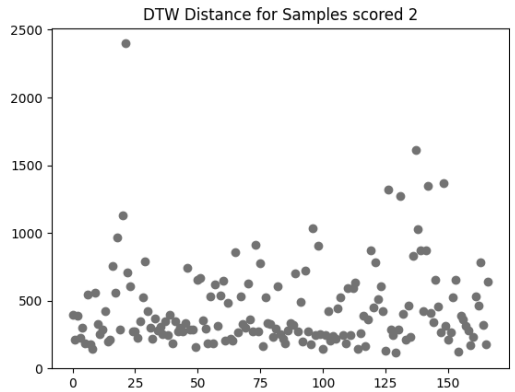

Figure 3: DTW Distance for Samples scored 1.

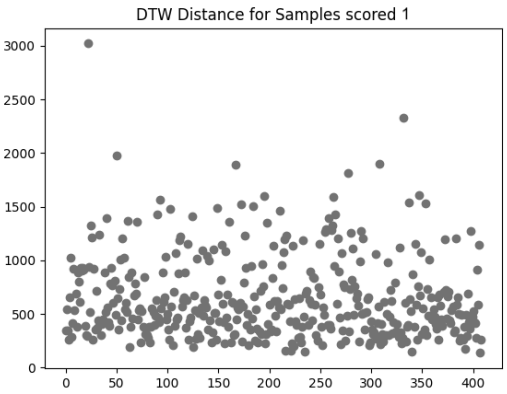

Figure 4: DTW Distance for Samples scored 2.

Table 3: Outlier Analysis using Median Absolute Deviation.

% of samples with Score 1 with $\boldsymbol{\lvert dist - mean \rvert > 3 * MAD}$	9.05%
% of samples with Score 2 with	10.18%

$\|dist - mean\| > 3 * MAD$	
% of samples with Score 3 with $\|dist - mean\| > 3 * MAD$	30.59%
% of samples with Score 4 with $\|dist - mean\| > 3 * MAD$	41.58%
Total % of samples with $\|dist - mean\| > 3 * MAD$	20.85%

```
Standard Deviation for score 4 samples: 317.7155764950189   Mean for score 4 samples: 308.65586532739934
Standard Deviation for score 3 samples: 324.7257357345005   Mean for score 3 samples: 375.7576950331108
STD for score 2 samples: 312.8142717850489   Mean for score 2 samples: 443.485981580863
STD for score 1 samples: 387.4253929974545   Mean for score 1 samples: 645.7864193917277

Number of Deviations considered for MAD: 2

Samples with Score of  1
    Number of Samples: 409
    MAD: 193.31531795105377
    Percentage of samples further than 2 MAD from the mean: 25.672371638141808
    Percentage of samples further than 3 MAD from the mean: 9.04645476772616

Samples with Score of  2
    Number of Samples: 167
    MAD: 118.68076918138357
    Percentage of samples further than 2 MAD from the mean: 29.34131736526946
    Percentage of samples further than 3 MAD from the mean: 10.179640718562874

Samples with Score of  3
    Number of Samples: 85
    MAD: 72.4236703646726
    Percentage of samples further than 2 MAD from the mean: 58.8235294117647
    Percentage of samples further than 3 MAD from the mean: 30.58823529411765

Samples with Score of  4
    Number of Samples: 279
    MAD: 56.48509653909062
    Percentage of samples further than 2 MAD from the mean: 75.62724014336918
    Percentage of samples further than 3 MAD from the mean: 41.577060931899645

Percentage of samples further than 2 MAD from the mean across all scores: 44.148936170212764
Percentage of samples further than 3 MAD from the mean across all scores: 20.851063829787233
```

Figure 2: Statistical Analysis of DTW Distance.

As shown in the Figure 5 and Table 3, 20.85% of samples across scores are outside of 3*MAD from the mean. This means that, if DTW distance is used to measure similarity, more than 20% of the samples are not judged accurately by humans. The number of outliers is much larger in the higher score samples, indicating that, when singing is poor, it is easy for humans to judge with consistency, but as the singing improves, only a trained ear can spot errors. This deduction is further proven by taking a few example performances and plotting performance and reference F0 data. Figure 6 displays a recording that looks similar to the reference yet was scored a 1 by the human graders. On the other hand, Figure 7 displays a recording that looks completely different from the reference yet was scored a 4 by humans. The DTW distance for these samples is also

given. As shown below, human scores are not aligned with the visual observation and the DTW distance at all.

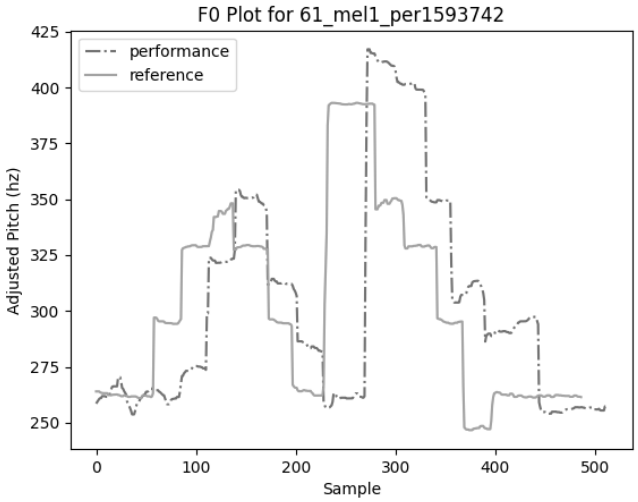

Figure 3: Human Scored 1, Visually looks 4, DTW Distance 188.11.

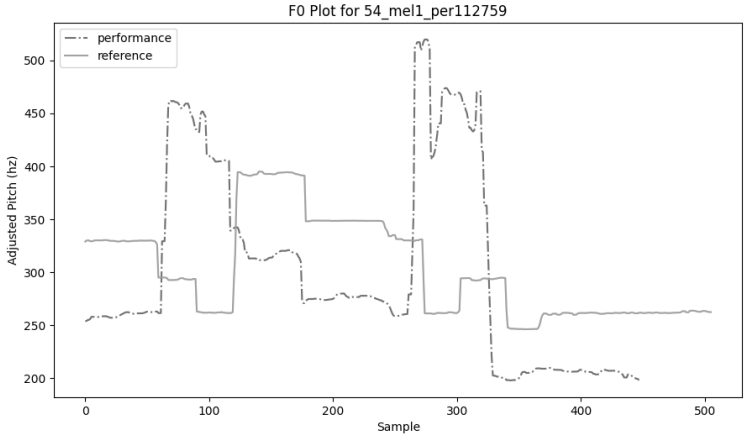

Figure 4: Human Scored 4, Visually looks 1, DTW Distance 1247.24.

6. Automated Holistic Scoring of Vocal Performance

Now that a deep understanding of our dataset and its inconsistencies has been established, I can train a model to predict scoring. I used the KNN (K-Nearest Neighbor) classification algorithm to assign the score to performance. The feature used for training was the DTW distance between performance and reference. Using the annotations from the original MAST Melody dataset, which contains the numerical scores of the performances, I created test and training datasets. Then, I use a KNN Classification model, with n=5 neighbors, to predict the scores for a test set. As noticed in the data analysis, there is a lot of outliers in the data when there is no consensus. The attempt to predict with all the samples in the dataset resulted in ~60% accuracy. This is quite low because of the inconsistencies shown in the dataset. Following this, two different experiments were conducted:

6.1. Experiment 1: Training and testing on all data (two categories)

Due to the large number of outliers and bias in the dataset (most samples scored 1 and 4), I conducted an experiment to attempt prediction on the entire dataset but only into two categories.

For this, the entire dataset was divided into two categories – category 1: scores 1 and 2 and category 2: scores 3 and 4. All the parameters for the KNN algorithm were the same as above (neighbors=5). The accuracy for this experiment using the same algorithm and parameter came to be 79.79%. This is a great outcome and implies that the model can predict a score for a performance with ~80% accuracy.

6.2. Experiment 2: Training and testing on consensus data

In the second experiment, I used only consensus samples (398) for training and testing. When using consensus samples, the accuracy of prediction was much better (> 80%). There were two sub-experiments conducted with consensus data:

1. 4 Category Prediction: Predict a score of 1-4 as available in the original annotation with only consensus data. The score reached with this approach was 80%.
2. 2 Category Prediction: As pointed out in the data analysis, there are very few consensus samples with scores 2 and 3, so I decided to reduce the data into two categories:
    a. samples with a score of 1 or 2
    b. and samples with a score of 3 or 4

    The goal is to classify all samples into two categories and then use that dataset for prediction. The confusion matrix (Table 4) below proves the same point and justifies the classification into two categories. The model was found to be 83.75% accurate with this approach.

Table 4 below shows the accuracy of prediction for each experiment. In addition the confusion matrix in Table 5 shows that most of the test samples were either scored 1 or 4 by humans.

Table 4: Prediction Accuracy.

4 Category Scoring Accuracy	80%
2 Category Scoring Accuracy	83.75%

Table 5: Confusion Matrix.

		Actual Score			
Predicted Score		1	2	3	4
	1	38	0	0	4
	2	1	0	0	1
	3	0	0	0	1
	4	9	0	0	26

7. Conclusion

The proposed AI model achieves greater 80% accuracy for predicting performances in the consensus set (the set of performances where human judges had consensus scoring). Many experiments were conducted to validate the model with different slices of data, and the accuracy was ~83% in some instances.

Table 6: Model Accuracy across experiments.

Data Set Used	Number of Categories	Accuracy
All samples used	4	60%
All Samples Used	2	79.89%
Consensus Samples Used	4	80%
Consensus Samples Used	2	83.75%

This implies that the AI model developed as part of this paper can score singing performances corresponding to a piece of reference music as well as a group of human experts in greater than 80% of cases.

In addition, the analysis of the F0 data of the vocal performance and references clearly shows that human grading is highly subjective. Table 9 demonstrates the point quantitatively. 20.85% of the samples were detected as outliers. This means that these samples are not judged accurately by humans. The number of outliers is much higher in highly scored samples, indicating that humans tend to ignore small mistakes made by the performer.

8. References

1. '"Dynamic time warping.' (2023)." Wikipedia, The Free Encyclopedia, Wikimedia Foundation, 30 August 2023,. Available at https://en.wikipedia.org/wiki/Dynamic_time_warping. (Accessed: 29 October 2023).
2. '"Fundamental frequency."' (2023). Wikipedia, The Free Encyclopedia, Wikimedia Foundation, 28 October 2023,. Available at https://en.wikipedia.org/wiki/Fundamental_frequency. (Accessed 29 October 2023).
3. B. Bozkurt, O. Baysal, D. Yuret (2017), 'A Dataset and Baseline System for Singing Voice Assessment', 13th Int. Symposium on Computer Music Multidisciplinary Research, https://api.semanticscholar.org/CorpusID:10732480.
4. Benetos, Emmanouil, Dixon, Simon, Duan, Zhiyao and Ewert, Sebastian (2019), 'Automatic Music Transcription: An Overview', IEEE Signal Processing Magazine, 36:1, https://doi.org/10.1109/MSP.2018.2869928.
5. Bozkurt, Baris and Baysal, Ozan (2023), MAST melody dataset, Zenodo, June 2023, https://doi.org/10.5281/zenodo.8007358.
6. Crawford, Renee (2013), 'Evolving Technologies Require Educational Policy Change: Music Education for the 21st Century', Australasian Journal of Educational Technology, 29:5, https://doi.org/10.14742/ajet.268.
7. Gao, Qiaozhan, PITCH DETECTION BASED MONOPHONIC PIANO TRANSCRIPTION, University of Rochester.
8. Gynnild, Vidar (2019), USING AN ASSESSMENT RUBRIC FOR FEEDBACK AND LEARNING: A CONCEPTUAL STUDY, 108-111, https://doi.org/10.36315/2019v1end023.
9. Hash, Philip M. (2021), "Supply and Demand: Music Teacher Shortage in the United States," *Research & Issues in Music Education*: Vol. 16: No. 1, Article 3. Available at: https://commons.lib.jmu.edu/rime/vol16/iss1/3
10. Klapuri, Anssi and Davy, Manuel (2006), Signal Processing Methods for Music Transcription, New York: Springer.
11. Liu, Chenchen (2023), 'Research advancement and foci of mobile technology-supported music education: a systematic review and social network analysis on 2008-2019 academic publications', Interactive Learning Environments, 31:7, https://doi.org/10.1080/10494820.2021.1974890.
12. Meert, W., Hendrickx, K., Van Craenendonck, T., Robberechts, P., Blockeel, H., & Davis, J. (2020). DTAIDistance (Version 2) [Computer software]. https://doi.org/10.5281/zenodo.3981067
13. Nart, Sevan (2016), 'Music Software in the Technology Integrated Music Education', TOJET: The Turkish Online Journal of Educational Technology, 15:2, https://files.eric.ed.gov/fulltext/EJ1096456.pdf, Accessed 14 October 2023.
14. Özer, Zeynep and Demirbatır, Rasim Erol (2023), 'Examination of STEAM-based Digital Didal Learning Applications in Music Education', European Journal of STEM Education, 8:1, https://doi.org/10.20897/ejsteme/12959.
15. Skiba, Richard (2020), 'Graded assessment models for competency-based training in vocational education and training', World Journal of Education, 10:3, https://doi.org/10.5430/wje.v10n3p106.
16. Sleep, Jonathan (2017), MA Thesis, San Luis Obispo: California Polytechnic State University.

17. Su, Li, Yang, Yi-Hsuan, (2016), Escaping from the Abyss of Manual Annotation: New Methodology of Building Polyphonic Datasets for Automatic Music Transcription, In: Kronland-Martinet, R., Aramaki, M., Ystad, S. (eds) Music, Mind, and Embodiment, CMMR 2015, Lecture Notes in Computer Science(), vol 9617, Springer, Cham, https://doi.org/10.1007/978-3-319-46282-0_20.
18. Yiting, Zhang and Sonquan, Yang (2022), 'Modern technology-enabled approaches in preschool music education', Interactive Learning Environments, 0:0, https://doi.org/10.1080/10494820.2022.2081211.
19. Zhao, Yanyan (2022), 'Analysis of Music Teaching in Basic Education Integrating Scientific Computing Visualization and Computer Music Technology', Mathematical Problems in Engineering, 2022, https://doi.org/10.1155/2022/3928889.

Author Biography

Saanvi Bhargava is a junior at The Harker School in San Jose. Saanvi loves to sing and has been performing since the age of 6. She started learning computer science in middle school and has been recognized for her problem-solving projects in national competitions and in the Synopsis Science Fair. She is the Founder and President of the Beats and Bytes Club, which explores how one can analyze music through technology. Saanvi has been researching how to make music education accessible to more people. Her learnings are published on her blog (https://medium.com/computational-musicology) and all her past projects can be found at (https://github.com/saanvib). A talented and trained singer in western classical, she is part of her school's prestigious show choir group as an alto. In addition to these interests, she is also President-elect for multiple clubs at her school most notably the Future Problem Solving club which focuses on solving practical problems facing our society. She is applying that interest, and her skill in coding and machine learning in exploring how to make music a part of more people's lives.

ORCID: https://orcid.org/0009-0007-3904-7753