Analysis of the Usefulness of Critique Documents on Musical Performance: Toward a Better Instructional Document Format

Masaki Matsubara$^{1[0000-0003-1950-683X]}$, Rina Kagawa$^{1[0000-0002-0482-5179]\star}$,
Takeshi Hirano[2], and Isao Tsuji[3,4]

[1] University of Tsukuba, Tsukuba, Japan masaki@slis.tsukuba.ac.jp
[2] University of Electro-Communications, Tokyo, Japan
[3] Senzoku Gakuen College of Music, Kawasaki, Japan
[4] Kunitachi College of Music, Tokyo, Japan

Abstract. Today, with the COVID-19 pandemic, the demand for remote and asynchronous lessons for musical performance is rapidly increasing. In these lessons, teachers listen to recordings of musical performances and then return textual critique documents to the performers. However, the common document formats that exist in other fields are not widely known in the field of performance instruction. To address this issue, we launched a project in 2020 to collect and publish a dataset of critique documents. This study describes a statistical analysis of the dataset to investigate which types of elements are useful for performers. The multilevel modeling results revealed that the content of the critiques differed more depending on the teacher than on the musical piece or the student. Particularly, the number of sentences about giving practice advice is a key factor for useful critique documents. These findings would lead to improved forms of critique documents and, eventually, to the development of educational programs for teachers.

Keywords: Textual Document Format· Digital Archive · Knowledge Management

1 Introduction

Instructions for playing musical instruments have traditionally been given face-to-face, and teaching in a virtual space has been considered unsuitable. However, today, with the coronavirus disease 2019 (COVID-19) pandemic, the demand for remote and asynchronous lessons is growing rapidly [1, 13]. For such lessons, teachers listen to recordings of musical performances write on textual critique documents, and then return them to performers. Today, some music colleges adopt this approach.

Remote and asynchronous lessons have an advantage in many aspects, such as facilitation of instruction to remote areas, flexible scheduling, and reduced

* Two authors equally contributed to this research.

travel. In particular, accumulated textual critique documents from the lessons have the potential for knowledge transfer and reuse; for example, students can use them for their practice, other students can also use them as references, and teachers can use them to improve their teaching methods. Although previous studies have focused on the transcription of speech in interactive instruction [2, 5–7, 9, 28, 29, 34], only a few have investigated textual documents for musical performance instruction. Hence, the format (i.e., description and arrangement) of textual documents for performance instruction [23] has not been clarified. In the digital era, the collection and utilization of textual critique documents for performance instruction are expected, but their reusability is low.

Therefore, as the first step in investigating useful document formats for performance education, this study aims to clarify what elements are effective for performers. Toward this goal, we launched a project for accumulating instructional textual data with people from a college of music in 2020. We have already collected and published a dataset that consists of 239 textual critique documents for 90 performances of 10 pieces by nine players [19, 20]. The analysis procedure was as follows. First, the usefulness of the critique documents was evaluated by the performers via crowdsourcing. Second, each sentence was categorized into six types by annotators. Third, multilevel modeling was conducted to clarify which types are useful.

The result with the best fitting model revealed the following two points. (1) Number of sentences that are describing practice strategy, objective information, feedback, and advice is effective for improving the usefulness of critique documents. (2) The usefulness of the critiques differed more depending on the teacher than the piece or the student. These findings can be applied to developing a improved document format by explicitly providing recommendations to input effective types of critique in the form.

2 Related Work

2.1 Music Database for Research

Several public datasets or digital archives have been constructed as knowledge resources for music with various perspectives[24], (e.g., performance recordings data [10], metadata [genre, composer, lyrics, etc. [11, 27, 30]], musical scores [MIDI [16], piano notation [8, 31]], information associated with [fingering [22], music analysis [12]], other multimodal information [18, 33], emotions [3, 35], listening history [26], and performers' interpretations [14, 15, 21, 25]). Most of the datasets are the data of the sound source itself. To the best of our knowledge, no datasets has shared human cognition, such as how they played or how they sounded.

2.2 Document Format

In some areas other than music, document formats have been designed. Regarding existing frameworks for document formats, IMRaD (introduction, method,

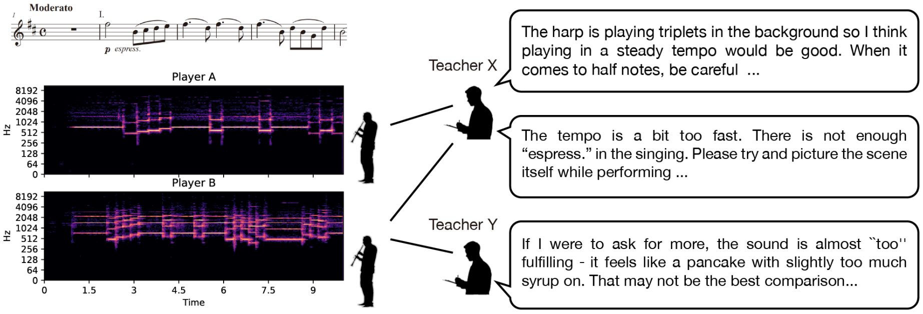
Moderato
I.
p espress.
Player A
8192
4096
2048
1024
512
256
128
64
0
Hz
Player B
8192
4096
2048
1024
512
256
128
64
0
Hz
0
1.5
3
4.5
6
7.5
9
Time
Teacher X
Teacher Y
The harp is playing triplets in the background so I think playing in a steady tempo would be good. When it comes to half notes, be careful ...
The tempo is a bit too fast. There is not enough "espress." in the singing. Please try and picture the scene itself while performing ...
If I were to ask for more, the sound is almost ``too'' fulfilling - it feels like a pancake with slightly too much syrup on. That may not be the best comparison...

Fig. 1. Examples of a CROCUS dataset [19, 20]. (Left) spectrogram of performance recordings of Tchaikovsky's piece by players A and B. (Right) Critique documents from teachers X and Y. From the observation, the tempo and the expression vary among players. Different teachers have different points of view.

result, and discussion) [4] has been a widely employed format for scientific writing. SOAP (subjective data, objective data, assessment, and plan) [32] is widely accepted worldwide in hospitals for clinical notes. In these research fields, document formats have been established, and document datasets are publicly available and actually utilized for research and education [17]. However, all of these are summaries of the opinions of experts in each field, and cannot be applied to performance education.

3 Materials and Methods

3.1 Materials

CROCUS dataset. In our project, we have already published a *CROCUS* (CRitique dOCUmentS of musical performance) dataset consisting 239 textual critique documents for performance instruction (in Japanese) for 90 performances of 10 pieces by nine players [19, 20]. Fig. 1 shows examples of the dataset.

Questionnaire survey of the usefulness of the documents. To evaluate the usefulness of the critique documents, we conducted an online questionnaire survey via a crowdsourcing platform. We recruited 200 people who had musical experience outside of school and asked them to provide their demographics and answer the question " Do you think that this document is useful for future performances? " with an 11-point Likert scale (10: useful – 0: useless). Each participant responded to 25 randomly selected textual documents for performance instruction.

Annotation. To clarify what types of element are effective for the usefulness of critique documents, annotators categorized every sentences into six types (Ta-

Table 1. Types of annotation for each sentence

Types	Example of sentence
Giving Subjective Information (GSI)	It is a very light and springy performance.
Giving Objective Information (GOI)	Tempo is late in the second bar.
Asking Question (AQ)	Is there a problem with the tuning of the instrument?
Giving Feedback (GF)	The pitch unconsciously moves during a vibrato.
Giving Practice (GP)	Please practice this phrase using the metronome.
Giving Advice (GA)	The first bar should have no crescendo.

ble 1). According to the Simones' definition [29], this study focused on the following six types: giving subjective information (GSI), giving objective information (GOI), asking question (AQ), giving feedback (GF), and giving advice (GA). Note that, giving information in Simones' definition is divided into GSI and GOI in this study. Other types such, as demonstrating, modeling, and listening/observing, were omitted because these cannot be implemented in remote and asynchronous teaching.

One of these six types was annotated for each sentence. Sentence breaks were periods or exclamation marks. When it was judged that one sentence consisted of descriptions of multiple types, it was separated by a comma. Each of the two annotators annotated for all 239 documents. If the annotations did not match, the final annotation was decided through discussion. The Cohen's Kappa coefficient was 0.96.

3.2 Method: Statistical Models for Analysis

Procedure. Multilevel modeling was conducted for the analysis. Multilevel modeling enables analysis assuming that the behavior of individual data changes depending on the hierarchy of data, that is, the group to which each data belongs. In other words, in this study, not only the change in usefulness among documents but also the influence of the teachers could be analyzed. We first tested the hierarchy of the characteristics of documents and then devised four models for analysis. R 4.1.0, brms 2.15.0, lme4 1.1–27, and lattice 0.20–38 were used.

Hierarchy of the usefulness. For teachers, the intraclass correlation coefficient (ICC) was 0.45, and the design effect (DE)[1] was 9.43. For players or pieces, ICCs were 0.0, and DEs were 1.0. Therefore, the usefulness scores showed hierarchy among teachers.

[1] DE is a criterion that takes into account both the average number of data in the group and ICC. A DE of over two suggests that the data are hierarchical. $DE = 1 + (k^* - 1)ICC$. k^* means the average number of data of group.

Models. Based on hierarchy, we devised the following four models:

Model I: The usefulness of the documents is affected by the presence or absence of each type.

Model II: The usefulness of the documents is affected by the number of descriptions of each type.

Model III: The usefulness of the documents is affected by the presence or absence of each type and varies depending on teachers.

Model IV: The usefulness of the documents is affected by the number of descriptions of each type and varies depending on teachers.

Let α be intercept, k be a content category, $\beta_k(k = 1, \ldots, 6)$ be coefficient of n_{ki}, n_{ki} be the number of descriptions for each type. In i-th document of the j-th participants, the usefulness of the k-th content is designated as follows:

$$\mathit{usefulness}_{ij} = \alpha + \sum_{k=1}^{6} \beta_k n_{jk} + \sum_{k=1}^{6} \eta_k^{(z_{ijk})} + \sum_{k=1}^{6} \gamma_k^{(z_{ijk})} n_{ik} + e_{ij}$$

Here, z_{ijk} indicates each teacher who wrote the i-th document. $\beta_k^{(z_{ijk})}$ is the random effect of the presence of unknown words on the intercept for the k-th content category of the i-th document. $\gamma_k^{(z_{ijk})}$ is the random effect of the presence of unknown words on the coefficient for n_{ik}.

The model parameters were fitted with four Markov chain Monte Carlo chains with 2,000 iterations and 1,000 burn-in samples with a thinning parameter of one. Non-informative priors were used for all estimations. Specifically, we used $\beta_k \sim N(0, 100)$, $\alpha \sim \mathit{StudentT}(3, 0, 2.5)$, and $\sigma_e \sim \mathit{StudentT}(3, 0, 2.5)$ as the prior distributions of the fixed effects, $\mathit{StudentT}(3, 0, 2.5)$ as the prior distribution of SD of random effects, and $\mathit{LKJCholesky}(1)$ as the prior distribution of the correlation matrix between $\gamma_k^{(g)}$ and $\eta_k^{(g)}$ for $k \in \{1, \ldots, 6\}$ and $g \in \{1, \ldots, 12\}$. The models were compared based on the widely applicable information criterion (WAIC). A smaller WAIC corresponds to a better model.

4 Results

4.1 Usefulness evaluation and types annotation

Fig. 2 shows the average scores of the usefulness for each teacher, student, and piece. This result implies that the usefulness of the critiques differed more depending on the teacher than on the piece or the student.

For the annotation results, the percentages of documents containing GSI, GOI, AQ, GF, GP, and GA were 47.28%, 54.81%, 3.34%, 39.33%, 22.18%, and 93.72%, and the average (and standard deviation) of the number of each category per document was, 0.70 (0.90), 0.85 (1.00), 0.03 (0.18), 0.61 (0.88), 0.33 (0.70), and 3.33 (2.50), respectively.

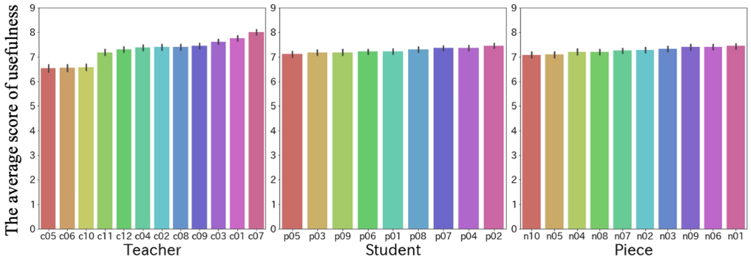

Fig. 2. Average scores of usefulness based on teacher, student, and piece.

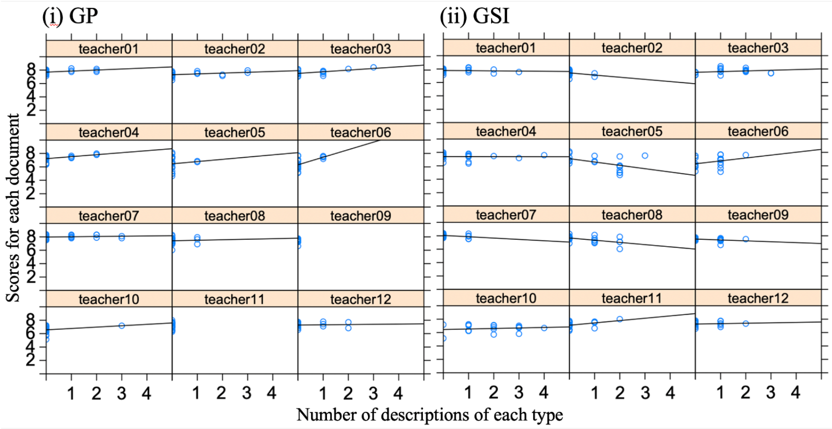

Fig. 3. Relationship between the number of description of GP and GSI in the document written by each of the 12 teachers and the usefulness score for each document. The relationship differed depending on the teachers. For example, the number of GP particularly increases the usefulness of the documents for teachers 06, 05, and 04 but not in teachers 07 and 12. The number of GSI decreased the usefulness in all teachers, except teachers 06 and 11, where the usefulness was increased.

4.2 Multilevel modeling

As fitting indices, the WAIC values for models I–IV were 449.6, 381.7, 320.0, and 292.1, respectively. All $\hat{R}$ were 1.01 or less. These results indicated that the model IV was the best model; that is, the number of descriptions of all types affects the usefulness, and these influences are affected by the teachers in all types. Table 2 shows the effect score of each type and teacher on usefulness. The results suggest that the more descriptions of GOI ($\beta_2 = 0.13, 95\%CI[0.05 - 0.20]$), GF ($\beta_4 = 0.13, 95\%CI[0.04 - 0.23]$), GP ($\beta_5 = 0.27, 95\%CI[0.09 - 0.46]$), and GA ($\beta_6 = 0.15, 95\%CI[0.07 - 0.22]$) significantly increased the usefulness, and GP had the highest usefulness among all the models.

Table 2. Effect of the number of descriptions of each type and teacher on document usefulness. GOI, GF, GP, and GA had positive lower-95% CI, among which GP showed the highest Estimate score. This indicates that usefulness will increase if the number of four types is increased, and that GP is the most effective type for instruction.

Population-Level Effects	Estimate	Est.Error	lower-95% CI	upper-95% CI	$\hat{R}$
Intercept	6.55	0.24	6.04	7.01	1.00
GSI	0.08	0.08	-0.08	0.23	1.00
GOI	0.13	0.04	0.05	0.20	1.00
AQ	0.54	0.92	-1.31	2.39	1.00
GF	0.13	0.05	0.04	0.23	1.00
GP	0.27	0.09	0.09	0.46	1.00
GA	0.15	0.04	0.07	0.22	1.00
Group-Level Effects					
sd(Intercept)	0.20	0.17	0.01	0.62	1.00
sd(GSI)	0.22	0.08	0.08	0.40	1.00
cor(Intercept,GSI)	-0.21	0.56	-0.97	0.91	1.01
sd(Intercept)	0.21	0.18	0.01	0.65	1.00
sd(GOI)	0.05	0.05	0.00	0.18	1.00
cor(Intercept,GOI)	-0.15	0.58	-0.97	0.93	1.00
sd(Intercept)	0.19	0.17	0.01	0.63	1.00
sd(AQ)	1.47	0.86	0.50	3.73	1.01
cor(Intercept,AQ)	-0.01	0.57	-0.94	0.95	1.00
sd(Intercept)	0.21	0.18	0.01	0.67	1.00
sd(GF)	0.08	0.07	0.00	0.25	1.01
cor(Intercept,GF)	-0.13	0.57	-0.97	0.92	1.00
sd(Intercept)	0.22	0.19	0.01	0.72	1.00
sd(GP)	0.21	0.11	0.03	0.48	1.00
cor(Intercept,GP)	-0.24	0.56	-0.97	0.89	1.00
sd(Intercept)	0.44	0.25	0.03	1.03	1.00
sd(GA)	0.10	0.04	0.03	0.19	1.00
cor(Intercept,GA)	-0.67	0.42	-1.00	0.62	1.01
Family Specific Parameters					
Sigma	0.40	0.02	0.36	0.45	1.00

Fig. 3 shows the relationship between the amount of each type of document written by each of the 12 teachers and the usefulness score for each document. The results showed that the relationship differs depending on teachers. For example, the number of GP generally increased the usefulness of the documents for teachers 06, 05, and 04 but not for teachers 07 and 12. The number of GSI decreased for all the teachers, except for teachers 06 and 11, where the usefulness was increased.

5 Discussion

The results of multilevel modeling analysis showed that the number of descriptions of all types tended to improve the usefulness of critique documents. In addition, a high number of descriptions of GA, GF, GOI, and GP significantly improved the usefulness. These findings would lead to better forms of critique documents where teachers are recommended to include GP, and GA, GF, and GOI if possible. We also confirmed that contents of the critiques differed more depending on the teacher than the piece or the student, suggesting that there is a need for supporting writing critique documents for teachers. To address this problem, considering how to express the sentence of the particular type is the future work.

In addition, the interaction between factors was not considered in the current study. For future work, we would like to examine the interaction between types (e.g., When descriptions of AG are numerous, the influence of GSI is small, and vise versa).

In the future, it is also necessary to investigate the effect of the player's knowledge or the relationship between the player and the teacher on usefulness should be investigated (e.g., whether a relationship of trust has been established, whether the relationship is still shallow).

6 Conclusion

This study showed that based on multilevel modeling, the number of descriptions of six types tended to improve the usefulness of the performance instruction document. Furthermore, the larger the number of descriptions of GA, GF, GOI, and GP, the more significant was the increase in the usefulness of the documents. The effect was different depending on the teacher. In the future, we would like to discuss the arrangement of documents, determine their format, and consider the development of educational programs and writing support technologies so that teachers can make these descriptions.

Acknowledgment

This study was partially supported by JST-Mirai Program Grant Number JP-MJMI19G8, and JSPS KAKENHI Grant Number JP19K19347 and JP21H03552.

References

1. Bayley, J.G., Waldron, J.: "it's never too late": Adult students and music learning in one online and offline convergent community music school. Int. J. Music. Educ. **38**(1), 36–51 (2020)
2. Cavitt, M.E.: A descriptive analysis of error correction in instrumental music rehearsals. J. Res. Music. Educ. **51**(3), 218–230 (2003)
3. Chen, Y.A., Yang, Y.H., Wang, J.C., Chen, H.: The amg1608 dataset for music emotion recognition. In: ICASSP. pp. 693–697 (2015)
4. Day, R.A., et al.: The origins of the scientific paper: the imrad format. J Am Med Writers Assoc **4**(2), 16–18 (1989)

5. Dickey, M.R.: A comparison of verbal instruction and nonverbal teacher-student modeling in instrumental ensembles. J. Res. Music. Educ. **39**(2), 132–142 (1991)
6. Duke, R.A.: Measures of instructional effectiveness in music research. Bull. Counc. Res. Music. Educ. pp. 1–48 (1999)
7. Duke, R.A., Simmons, A.L.: The nature of expertise: Narrative descriptions of 19 common elements observed in the lessons of three renowned artist-teachers. Bull. Counc. Res. Music. Educ. pp. 7–19 (2006)
8. Foscarin, F., McLeod, A., Rigaux, P., Jacquemard, F., Sakai, M.: ASAP: a dataset of aligned scores and performances for piano transcription. In: ISMIR. pp. 534–541 (2020)
9. Goolsby, T.W.: Verbal instruction in instrumental rehearsals: A comparison of three career levels and preservice teachers. J. Res. Music. Educ. **45**(1), 21–40 (1997)
10. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: Rwc music database: Popular, classical and jazz music databases. In: ISMIR. pp. 287–288 (2002)
11. Goto, M., Hashiguchi, H., Nishimura, T., Oka, R.: Rwc music database: Music genre database and musical instrument sound database. In: ISMIR. pp. 229–230 (2003)
12. Hamanaka, M., Hirata, K., Tojo, S.: Gttm database and manual time-span tree generation tool. In: SMC. pp. 462–467 (2018)
13. Hash, P.M.: Remote learning in school bands during the covid-19 shutdown. J. Res. Music. Educ. **68**(4), 381–397 (2021)
14. Hashida, M., Matsui, T., Katayose, H.: A new music database describing deviation information of performance expressions. In: ISMIR. pp. 489–494 (2008)
15. Hashida, M., Nakamura, E., Katayose, H.: Constructing pedb 2nd edition: a music performance database with phrase information. In: SMC. pp. 359–364 (2017)
16. Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.Z.A., Dieleman, S., Elsen, E., Engel, J., Eck, D.: Enabling factorized piano music modeling and generation with the MAESTRO dataset. In: ICLR (2019)
17. Kagawa, R., Baba, Y., Tsurushima, H.: Publicly available medical text data with authentic quality (Oct 2020). https://doi.org/10.5281/zenodo.4064153, https://doi.org/10.5281/zenodo.4064153
18. Li, B., Liu, X., Dinesh, K., Duan, Z., Sharma, G.: Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. IEEE Tran. Multimedia **21**(2), 522–535 (2018)
19. Matsubara, M.: Crocus: Dataset of musical performance critique (Jun 2021). https://doi.org/10.5281/zenodo.4748243
20. Matsubara, M., Kagawa, R., Hirano, T., Tsuji, I.: Crocus: Dataset of musical performance critiques: Relationship between critique content and its utility. In: CMMR (2021)
21. Miragaia, R., Reis, G., de Vega, F.F., Chávez, F.: Multi pitch estimation of piano music using cartesian genetic programming with spectral harmonic mask. In: 2020 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1800–1807. IEEE (2020)
22. Nakamura, E., Saito, Y., Yoshii, K.: Statistical learning and estimation of piano fingering. Information Sciences **517**, 68–85 (2020)
23. Reiter, E., .D.R.: Building Natural Language Generation Systems. Cambridge: Cambridge University Press (2000)
24. Salamon, J.: What's broken in music informatics research? three uncomfortable statements. In: 36th International Conference on Machine Learning (ICML), Workshop on Machine Learning for Music Discovery. Long Beach, CA, USA (2019)

25. Sapp, C.S.: Comparative analysis of multiple musical performances. In: ISMIR. pp. 497–500 (2007)
26. Schedl, M.: The lfm-1b dataset for music retrieval and recommendation. In: ICMR. pp. 103–110 (2016)
27. Silla Jr, C.N., Koerich, A.L., Kaestner, C.A.: The latin music database. In: ISMIR. pp. 451–456 (2008)
28. Simones, L., Schroeder, F., Rodger, M.: Categorizations of physical gesture in piano teaching: A preliminary enquiry. Psychology of Music **43**(1), 103–121 (2015)
29. Simones, L.L., Rodger, M., Schroeder, F.: Communicating musical knowledge through gesture: Piano teachers' gestural behaviours across different levels of student proficiency. Psychology of Music **43**(5), 723–735 (2015)
30. Sturm, B.L.: An analysis of the gtzan music genre dataset. In: ACM Workshop MIRUM. pp. 7–12. MIRUM '12 (2012)
31. Wang, Z., Chen, K., Jiang, J., Zhang, Y., Xu, M., Dai, S., Bin, G., Xia, G.: Pop909: A pop-song dataset for music arrangement generation. In: ISMIR (2020)
32. Weed, L.L.: Medical records, medical education, and patient care: the problem-oriented record as a basic tool. Press of Case Western Reserve University (1969)
33. Weiß, C., Zalkow, F., Arifi-Müller, V., Müller, M., Koops, H.V., Volk, A., Grohganz, H.G.: Schubert winterreise dataset: A multimodal scenario for music analysis. J. Comp. Cult. Herit. **14**(2), 1–18 (2021)
34. Whitaker, J.A.: High school band students' and directors' perceptions of verbal and nonverbal teaching behaviors. J. Res. Music. Educ. **59**(3), 290–309 (2011)
35. Zhang, K., Zhang, H., Li, S., Yang, C., Sun, L.: The pmemo dataset for music emotion recognition. In: ICMR. pp. 135–142 (2018)