# CROCUS: Dataset of Musical Performance Critiques
## Relationship between Critique Content and Its Utility

Masaki Matsubara[1*], Rina Kagawa[1*], Takeshi Hirano[2], and Isao Tusji[3]

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

**Abstract.** In musical performance education, verbal as well as non-verbal information is used to convey knowledge. In the current social situation, the demand for remote and asynchronous lesson is increasing, and it is not clear what types of verbal information should be used. In this study, we collected 239 critique documents written in Japanese by 12 teachers for 90 performances of the same 10 orchestra studies of the oboe by 9 students. We categorized the critiques and found that their content differed more by the teacher than by the piece or the student. We also found that the category of *giving a practice strategy* was particularly valued by students.

**Keywords:** Database, Music Education, Verbal Information

## 1 Introduction

Playing musical instruments has traditionally been taught in-person and was considered unsuitable for virtual learning environments. However, the COVID-19 pandemic has led to increased demand for online music tuition [1, 17]. In musical performance tuition, knowledge is conveyed using both non-verbal information such as singing melodies and gesturing, and verbal information such as pointing out mistakes [7, 12, 22]. Verbal information is essential for conveying how the learner's performance sounds, why they did not play well, and how they should practice. An advantage of online music tuition is that space and time do not necessarily have to be shared, thus allowing for remote and asynchronous teaching. However, due to the low resolution of online video/audio communication, there is a limitation in the use of non-verbal information, as it is difficult to convey complex body movements and high-quality sound performances. Therefore, the importance of verbal information in music tuition, especially in the critique documents of asynchronous online performance education, is expected to increase [10].

However, teaching using words is not easy. In our preliminary survey of nine music college students and one hundred people with musical performance experience, they reported a good impression of their musical experience, although some were not satisfied with their teacher's instructions. We collected free-text responses about

[*] Two authors equally contributed to this research.

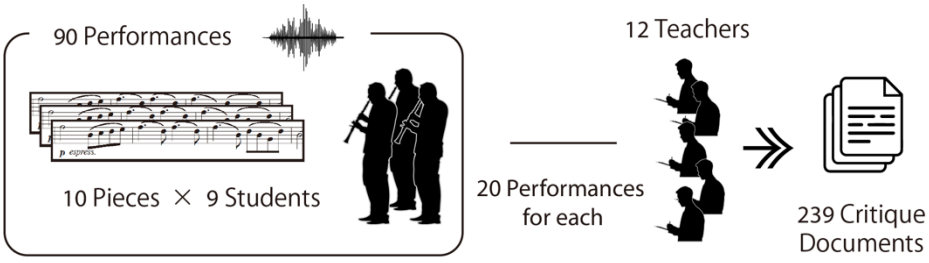


**Fig. 1.** Overview of the construction of the musical performance critique dataset.

dissatisfaction with the instruction and categorized the results as the pertaining to three issues: (1) content of performance instruction (e.g., "I would have preferred instruction based on facts," "lack of concrete advice"), (2) consistency of instruction over multiple lessons (e.g., "completely different or inconsistent attention from lesson to lesson"), and (3) wording of instruction not related to performance (e.g., "all the teacher did was scold without much praise").

We believe that one reasons for these problems is the lack of teaching protocols in performance instruction and the lack of systematic clarification of what should be verbalized to benefit learners. At present, however, the empirical knowledge of what types of instruction are provided is not widely shared, even among students who aspire to become professionals.

This paper introduces an open dataset of musical performance critiques in Japanese, called *CROCUS* (CRitique dOCUmentS of musical performance), to promote music education through the study of verbal information in performance instruction. We define critique documents as comments written by teachers to give feedback on a performance. We collected 239 critique documents from 12 teachers for 90 performances of 10 pieces by 9 students (Fig. 1). Because music college classes are conducted online at present, we collected recordings and critique documents similar in manner to those in asynchronous classes. This dataset allows us to compare critiques for each piece, student, and teacher. We examined which critique contents were perceived by the performers as useful instruction. Specifically, we analyzed types of verbal information, measured the perceived utility of critiques, and examined differences in utility scores among teachers, students, and pieces.

The contributions of this paper are as follows:

- We constructed an open dataset of 239 critique documents of 90 musical performances of 10 oboe orchestral studies[1].
- We quantitatively demonstrated that the content of the critique documents varies more by teacher than by piece or student.
- We collected evaluations of the critique documents from people with musical experience and examined the types of verbal information to determine what in the critique documents was described as having high utility.

[1] Dataset is public on https://doi.org/10.5281/zenodo.4748243

**Table 1.** List of Pieces

| ID | Composer | Piece |
|---|---|---|
| p01 | L. v. Beethoven | Symphony No. 3 in E flat Major 'Eroica', Op. 55 |
| p02 | G. A. Rossini | 'La Scala di seta' Overture |
| p03 | F. Schubert | Symphony No. 8 in B Minor D.759 'Unfinished' |
| p04 | J. Brahms | Violin Concerto in D Major, Op. 77 |
| p05 | P. I. Tchaikovsky | Symphony No. 4 in F minor, Op. 36 |
| p06 | P. I. Tchaikovsky | "Swan Lake", Ballet Suite, Op.20a |
| p07 | N. Rimsky-Korsakov | "Scheherazade", Symphonic Suite, Op. 35 |
| p08 | R. Strauss | "Don Juan", Symphonic Poem, Op. 20 |
| p09 | M. Ravel | Le Tombeau de Couperin I.Prelude |
| p10 | S. Prokofiev | "Peter and the Wolf", Symphonic Tale, Op. 67 |

## 2 Related Work

### 2.1 Music Database for Research

Numerous music databases have been published, with, for example, performance recordings data [14], metadata (genre, composer, lyrics, etc. [15, 28, 32]), musical scores (MIDI [20], piano notation [11]), information associated with scores (fingering [25], music analysis [16]), and other multimodal information [23]. There are also databases about the human aspects involved in music, including emotions [5], listening history [27], and performer interpretations [18, 19, 26], but, to the best of our knowledge, no database shares critiques in performance education.

### 2.2 Teaching Behavior in Musical Performance Tuition

Teaching behavior in musical performance tuition has been studied in the music education field, including comparison of teacher levels [13], analysis on time allocation [4], comparison [33] and categorization [29, 30] of verbal and non-verbal information, and teacher–student interaction [9]. These studies targeted the transcription of speech in interactive instruction. Our study focused on critique documents that can be used for asynchronous education.

Regarding the utility of instruction, one study compared verbal and non-verbal instruction [6], and another summarized the evaluation of its usefulness [8]. These studies were based on five or fewer performances. We conducted a large-scale study and clarified the relationship between the verbal information and its utility.

## 3 Method

We constructed the CROCUS dataset by collecting performance recordings and critique documents. Then, all comments in the documents were annotated. Finally, the perceived utility of every document was evaluated. All collection procedures were approved by the ethical review boards of the University of Tsukuba, Senzoku Gakuen College of Music, and Kunitachi College of Music.

**Table 2.** Types of verbal information in this study

| | Type | Definition |
|---|---|---|
| Adopted and adapted from Carlin [3]; Zhukov [34]; Simones [30] | Giving Subjective Information (GSI)[2] | Providing general and/or specific conceptual information based on the teacher's subjectivity. |
| | Giving Objective Information (GOI)[2] | Providing general and/or specific conceptual information based on objectively referable events or concepts. |
| | Asking Question (AQ) | Enquiring. |
| | Giving Feedback (GF) | Evaluation of a student's applied and/or conceptual knowledge. |
| | Giving Practice (GP) Strategy | Providing suggestions for ways to practice a particular passage or discussing a practice schedule. |
| | Giving Advice (GA) | Giving a specific opinion or recommendation to guide the student's action toward the achievement of specific musical aims, without demonstration or modeling. |

### 3.1 Constructing the CROCUS Dataset

A total of 90 performances (10 pieces by 9 music college students majoring in the oboe) were recorded. As online lessons have become the norm in the music colleges due to COVID-19, we adopted a comparable situation. Each student played in a less reverberant and less noisy environment at home, about 1 m away from the recording device (Roland R-07). Tuning and recording level were adjusted at the beginning of the recording. We selected the 10 pieces in Table 1 to balance difficulty, style, form, and era.

### 3.2 Annotating Types of Commentary in CROCUS

We adopted and adapted Simones' definitions [30], as shown in the Table 2[3]. One of these six types was annotated to each sentence. Sentence breaks were periods or exclamation marks. When a sentence was judged as consisting of multiple types, they were separated by a comma. Two annotators annotated all 239 documents. If the annotations did not match, the final annotation was decided through discussion. The Cohen's kappa coefficient was 0.96.

### 3.3 Evaluating Perceived Utility of Critique Documents

The perceived utility of the collected critiques was examined by 200 people who answered the question "Do you think that this document is useful for future performances?" by using an 11-point Lickert-like scale (10: useful – 0: useless). Participants responded to 25 randomly selected critique documents. This question is referred to as Q1.

[2] Originally "Giving Information." Divided by the authors.

[3] Types of "Demonstrating", "Modelling", and "Listening/Observing" were omitted because these actions are not observed in a written text.

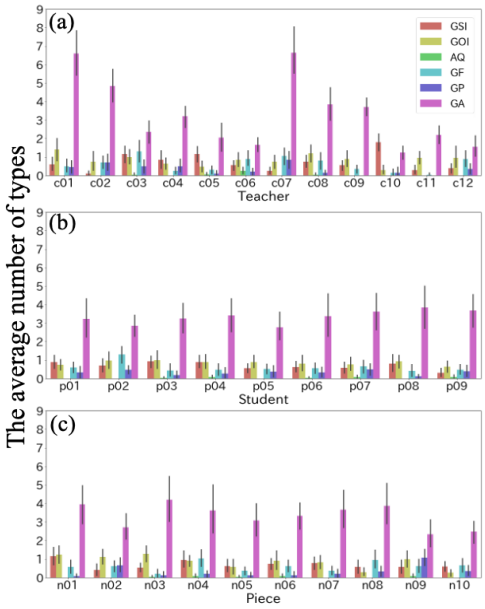


**Fig. 2.** Number of commentary types for each document per (a) teacher, (b) student, and (c) piece. The error bar indicates the standard deviation.

**Detailed Analysis of the Utility:** Utility is not limited to usefulness for future performances. Therefore, we referred to the usefulness perspective used in software requirement specifications [21] and accounting documents [31] for eight other items. The questions we used were as follows. All questions were asked in the form of "Do you think that this document —?" Q2: is readable, Q3: is understandable, Q4: has language not related to future performances, Q5: is not ambiguous, Q6: contains only statements related to the future performances, Q7: is consistent, Q8: can be verified by listening to the performance, and Q9: allows you to refer to the relevant part in the score from the content described.

## 4 Results

### 4.1 Constructing the CROCUS dataset

A total of 239 critique documents[4] were provided by 12 teachers who are currently with or formerly belonged to well-known music colleges, orchestras, and brass bands in Japan. Each teacher wrote critique documents assuming the usual lessons for a total of 20 performance recordings. The 20 performances were selected in a counterbalanced manner with the following constraints: each piece was reviewed by at least two teachers, and every teacher reviewed at least one performance for every student. Due to the current social situation, the critique documents were also written at the teacher's home.

[4] One critique was lost during the collection process.

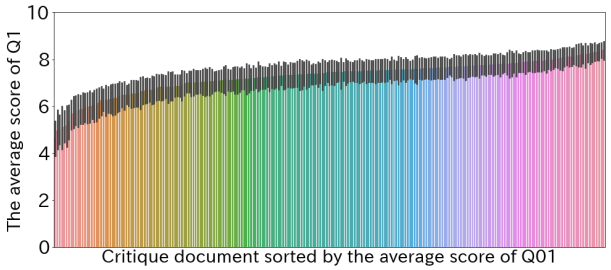


**Fig. 3.** Average score of Q1 for each critique document (sorted by Q1 score).

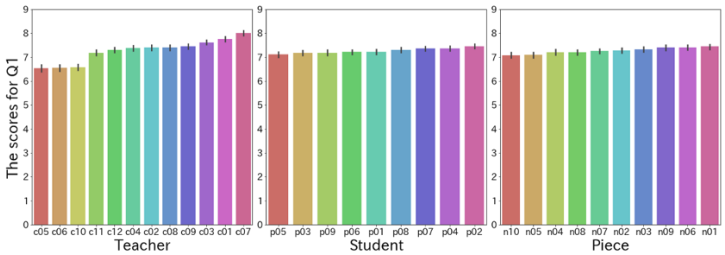


**Fig. 4.** Average scores for Q1 by teacher, student, and piece (sorted by Q1 score).

An example of a critique document is as follows:

*The difficult passages are performed well here. 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...*

### 4.2 Annotating Types of Commentary in CROCUS

GSI, GOI, AQ, GF, GP, and GA appeared in 47.28%, 54.81%, 3.34%, 39.33%, 22.18%, and 93.72% of the documents, respectively. The average (and standard deviation) of the number for 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 shows that the differences in the content of documents were larger among teachers than among songs or students.

### 4.3 Evaluating Perceived Utility of Critique Documents

Our results showed that the critique documents had a variety of utility scores, and there were documents that the readers perceived as less useful (Fig. 3). Since the null hypothesis that the distributions of Q1 values for each teacher, student, and piece (Fig. 4) were normal was rejected by the Shapiro-Wilk test, the Kruskal-Wallis test was used. The null hypothesis that the Q1 values that the reader perceived useful were equal among all teachers and the null hypothesis that they were equal among all pieces were rejected ($p \leq 0.001$, the effect size was small).

**Table 3.** The p-value and effect size of the Kruskal–Wallis test conducted for each question item for each teacher, student, and piece.

| | Q1 | Q2 | Q3 | Q4 | Q5 | Q6 | Q7 | Q8 | Q9 |
|---|---|---|---|---|---|---|---|---|---|
| Teacher | ***(s) | ***(m) | ***(s) | ***(s) | ***(m) | ***(m) | ***(s) | ***(s) | ***(m) |
| Student | **(vs) | ***(vs) | ***(vs) | ***(vs) | ***(vs) | ***(vs) | *(vs) | ***(vs) | ***(vs) |
| Piece | ***(vs) | ***(vs) | − (vs) | − (vs) | ***(vs) | ***(vs) | *(vs) | ***(vs) | ***(s) |

* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$; m, s, and vs mean moderate, small, and very small, respectively.

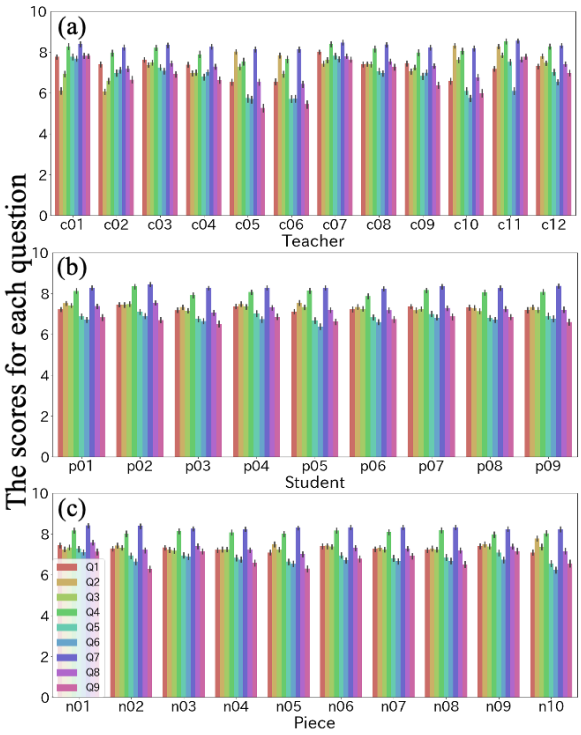


**Fig. 5.** Question scores for each document per teacher, student, and piece.

**Detailed Analysis of the Utility:** Table 3 presents the results of the Kruskal–Wallis test performed for the average score of each question (Fig. 5). The difference between the teachers was more remarkable than that between the students or the pieces for all questions. This result was consistent with previous research showing that the usage of words differs depending on the teachers [22]. The differences among the teachers were particularly large regarding whether the commentary was easy to understand (Q3), not ambiguous (Q5), contained only descriptions related to future performances (Q6), and can refer to the relevant part in the score (Q9). Detailed statistical analysis of the utility score of critiques is described in [24].

The critique documents with the highest average Q1 and that with the lowest average value are shown as follows (types are annotated with square bracket):

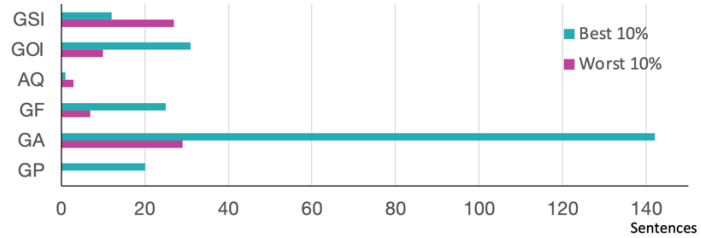


**Fig. 6.** Histogram of types in best 10% and worst 10% of critique documents.

**The highest-rated critique (n09-p04-c03) (Q1: 8.41 ± 1.44)**
[GSI] *I feel that this performance is very good, and it leaves a very favorable impression.* [GA] *Because of this, I would like you to be a little more careful regarding the nuances of the performance.* [GP] *Please practice the grace notes in bars 2 and 4 again by themselves. The same for bar 10.* [GF] *There is always a mistake in the E♭-H transition in bar 11.* [AQ] *Perhaps it is a problem with the tuning of the instrument?* [GP] *Please perform this part slowly and check carefully.* [GF] *If it is not a tuning problem, then I believe it is a fingering or breathing problem.* [GP] *Please practice carefully and check if the breathing and fingering are both coordinated properly.* [GA] *In the second half, there is tenuto on the high E and D notes.* [GA] *Please endeavor to perform each note carefully with nuance.*

**The lowest-rated critique (n04-p06-c05) (Q1: 4.63 ± 2.61)**
[GSI] *The melodies are performed beautifully and vibrantly, almost as if I could hear an orchestra performing.* [GSI] *The phrasings are well expressed for the piece, and it was lovely.*

## 5 Discussion

As Fig. 3, Fig. 4, and Table 3 indicate, the utility scores of critiques differed more greatly than by piece or each student. Teachers whose comments received high utility scores (e.g., c07, c01, and c03) provided information on the performance based on objective evidence (GOI), indicated the direction that the student should aim for (GA), and suggested practice strategies (GP). Fig. 6 is a histogram of types in the best 10% and worst 10% critique documents in terms of Q1 score. Giving practice strategies (GP) were observed in the best 10% and not in the worst 10%. Giving advice (GA), providing information based on objective facts (GOI), and giving feedback (GF) were also more common in highest-rated critiques.

## 6 Conclusion

We published the CROCUS dataset as a starting point for investigating the use of language in critique documents. The dataset clarified that the contents of critiques varied most by teacher, and suggested that the category of giving a practice strategy was valued by students.

Since the dataset was constructed at an early stage of the project, the instrument was limited to the oboe. Whether the findings are generalizable remains as an open question. We would like to explore more instruments and discuss whether a good critique document structure has a common characteristic. The student participants were limited to music college students; thus, we would like to explore the topic at various levels of experience, such as professional and amateur students. Finally, the study was conducted using only Japanese. In the future, it will be necessary to conduct comparisons among multiple languages and discuss differences between languages and cultures [2].

## Acknowledgments

This study was partially supported by JST-Mirai Program Grant Number JPMJMI19G8, and JSPS KAKENHI Grant Number JP19K19347. We would like to thank all the performers and teachers who participated in the data collection. We would also thank to those who helped us with data annotation and evaluation.

## References

1. Bayley, J. G. and 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., Vol. 38, No. 1, pp. 36–51 (2020).
2. Campbell, P. S.: Lessons from the world: A cross-cultural guide to music teaching and learning, MacMillan Publishing Company (1991).
3. Carlin, K. D.: Piano pedagogue perception of teaching effectiveness in the preadolescent elementary level applied piano lesson as a function of teacher behavior, PhD Thesis, Indiana University (1997).
4. Cavitt, M. E.: A descriptive analysis of error correction in instrumental music rehearsals, J. Res. Music. Educ., Vol. 51, No. 3, pp. 218–230 (2003).
5. Chen, Y.-A., Yang, Y.-H., Wang, J.-C. and Chen, H.: The AMG1608 dataset for music emotion recognition, ICASSP, pp. 693–697 (2015).
6. Dickey, M. R.: A comparison of verbal instruction and nonverbal teacher-student modeling in instrumental ensembles, J. Res. Music. Educ., Vol. 39, No. 2, pp. 132–142 (1991).
7. Dorman, P. E.: A review of research on observational systems in the analysis of music teaching, Bull. Counc. Res. Music. Educ., pp. 35–44 (1978).
8. Duke, R. A.: Measures of instructional effectiveness in music research, Bull. Counc. Res. Music. Educ., pp. 1–48 (1999).
9. Duke, R. A. and 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).
10. Dye, K.: Student and instructor behaviors in online music lessons: An exploratory study, Int. J. Music. Educ., Vol. 34, No. 2, pp. 161–170 (2016).
11. Foscarin, F., McLeod, A., Rigaux, P., Jacquemard, F. and Sakai, M.: ASAP: a dataset of aligned scores and performances for piano transcription, ISMIR, pp. 534–541 (2020).
12. Froehlich, H.: Measurement dependability in the systematic observation of music instruction: A review, some questions, and possibilities for a (new?) approach., Psychomusicology, Vol. 14, No. 1-2, p. 182 (1995).

13. Goolsby, T. W.: Verbal instruction in instrumental rehearsals: A comparison of three career levels and preservice teachers, J. Res. Music. Educ., Vol. 45, No. 1, pp. 21–40 (1997).
14. Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical and Jazz Music Databases., ISMIR, pp. 287–288 (2002).
15. Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Music genre database and musical instrument sound database, ISMIR, pp. 229–230 (2003).
16. Hamanaka, M., Hirata, K. and Tojo, S.: GTTM database and manual time-span tree generation tool, SMC, pp. 462–467 (2018).
17. Hash, P. M.: Remote learning in school bands during the COVID-19 shutdown, J. Res. Music. Educ., Vol. 68, No. 4, pp. 381–397 (2021).
18. Hashida, M., Matsui, T. and Katayose, H.: A New Music Database Describing Deviation Information of Performance Expressions., ISMIR, pp. 489–494 (2008).
19. Hashida, M., Nakamura, E. and Katayose, H.: Constructing PEDB 2nd Edition: a music performance database with phrase information, SMC, pp. 359–364 (2017).
20. Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., Elsen, E., Engel, J. and Eck, D.: Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset, ICLR (2019).
21. IEEE: Recommended Practice for Software Requirements Specifications, IEEE Std 830-1998, pp. 1–40 (1998).
22. Lehmann, A. C., Sloboda, J. A., Woody, R. H., Woody, R. H. et al.: Psychology for musicians: Understanding and acquiring the skills, Oxford University Press (2007).
23. Li, B., Liu, X., Dinesh, K., Duan, Z. and Sharma, G.: Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications, IEEE Tran. Multimedia, Vol. 21, No. 2, pp. 522–535 (2018).
24. Matsubara, M., Kagawa, R., Hirano, T. and Tsuji, I.: Analysis of Usefulness of Critique Documents on Musical Performance: Toward better Instructional Document Format, ICADL (to appear) (2021).
25. Nakamura, E., Saito, Y. and Yoshii, K.: Statistical learning and estimation of piano fingering, Information Sciences, Vol. 517, pp. 68–85 (2020).
26. Sapp, C. S.: Comparative Analysis of Multiple Musical Performances., ISMIR, pp. 497–500 (2007).
27. Schedl, M.: The LFM-1b Dataset for Music Retrieval and Recommendation, ICMR, pp. 103–110 (2016).
28. Silla Jr, C. N., Koerich, A. L. and Kaestner, C. A.: The Latin Music Database, ISMIR, pp. 451–456 (2008).
29. Simones, L., Schroeder, F. and Rodger, M.: Categorizations of physical gesture in piano teaching: A preliminary enquiry, Psychology of Music, Vol. 43, No. 1, pp. 103–121 (2015).
30. Simones, L. L., Rodger, M. and Schroeder, F.: Communicating musical knowledge through-gesture: Piano teachers' gestural behaviours across different levels of student proficiency, Psychology of Music, Vol. 43, No. 5, pp. 723–735 (2015).
31. Smith, M. and Taffler, R.: Readability and understandability: Different measures of the textual complexity of accounting narrative, Accounting, Auditing & Accountability Journal, Vol. 5, No. 4, pp. 84–98 (1992).
32. Sturm, B. L.: An Analysis of the GTZAN Music Genre Dataset, ACM Workshop MIRUM, MIRUM '12, pp. 7–12 (2012).
33. Whitaker, J. A.: High school band students' and directors' perceptions of verbal and nonverbal teaching behaviors, J. Res. Music. Educ., Vol. 59, No. 3, pp. 290–309 (2011).
34. Zhukov, K.: Teaching styles and student behaviour in instrumental music lessons in Australian conservatoriums, PhD Thesis, University of New South Wales (2005).