STOCHASTIC MODELING OF A MUSICAL PERFORMANCE WITH EXPRESSIVE REPRESENTATIONS FROM THE MUSICAL SCORE

Kenta Okumura, Shinji Sako and Tadashi Kitamura
Nagoya Institute of Technology, Japan
{k09,sako,kitamura}@mmsp.nitech.ac.jp

ABSTRACT

This paper presents a method for describing the characteristics of human musical performance. We consider the problem of building models that express the ways in which deviations from a strict interpretations of the score occurs in the performance, and that cluster these deviations automatically. The clustering process is performed using expressive representations unambiguously notated on the musical score, without any arbitrariness by the human observer. The result of clustering is obtained as hierarchical tree structures for each deviational factor that occurred during the operation of the instrument. This structure represents an approximation of the performer's interpretation with information notated on the score they used during the performance.

This model represents the conditions that generate the difference in the fluctuation of performance expression and the amounts of deviational factors directly from the data of real performance. Through validations of applying the method to the data measured from real performances, we show that the use of information regarding expressive representation on the musical score enables the efficient estimation of generative-model for the musical performance.

1. INTRODUCTION

The idea of having a computer perform like human musician arose more than two decades ago. There have been various proposals for making a computer understand the rich expression of a performance [2]. Historically, the mainstream approach to capturing the nuances of performance has changed from rule-based methods to learning-based methods. One model that shows the effectiveness of the latter approach is represented by the generative model. Also, there is another motivation for this kind of research, that is, learning what makes a performance humanlike; however, there are few initiatives based on such questions. One approach to analyze performance statistically, by capturing the trends of the performance in the acoustic features, has already been attempted [3, 8, 10, 11]. These studies are admirable in that their verification used a large quantity of expressive performance; we also essentially agree that it is desirable to perform the verification with such an approach. However, it is difficult to observe the expressiveness of a performance from diverse perspectives by these approaches as expressiveness consists of various factors. We adopt a MIDI-based approach to simplify such problems, and consider a variety of expressive representations notated on the musical score as the factor that describes how the expressive performance has been generated. In addition, our method to capture the performance is based on the idea of a generative model. Therefore, our method has the potential to generate an unseen performance, not merely to analyze an already known one.

In the following sections, we propose a method for the automatic analysis of the characteristics of a performance based on various combinations of expressive representations. Also, we observe what kinds of representation constitute the human quality of the performance by apply them to the data measured from the real performance to evaluate the validity of this method.

2. METHOD

In this section, we propose a method for the automatic classification of trends of the deviations in performance, so as to describe the dependencies between score and performance. On the keyboard instrument, a performer's key operation, in terms of timing and intensity, causes deviations from the score for the purpose of artistic expression. We believe that the performer's individuality would occur in the differences in the trend of deviations. The occurrence tendencies of these deviations in the performance are not constant, as they are affected by various factors such as the differences in musical compositions. To capture the characteristics of individuals who performed only in terms of deviation from the average trend in the overall performance is difficult; therefore, it is necessary to handle deviations in each key action, specifically and in general. Using this awareness, we have been studying a method that regards the trends in the deviation as a stochastic model and acquire these trends via learning and instructions on the score.

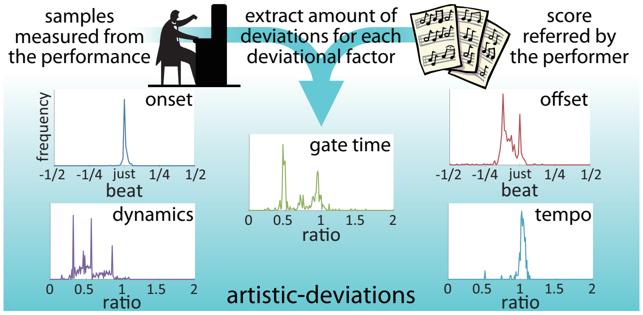


Figure 1. Extraction of deviational factors

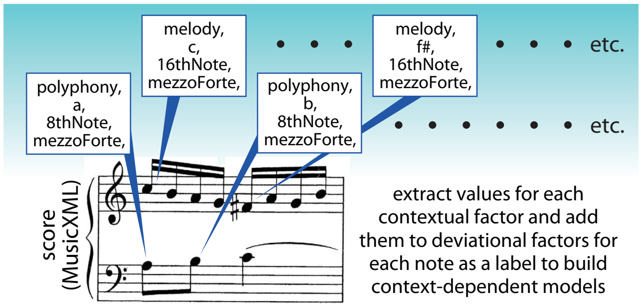

polyphony, a, 8thNote, mezzoForte,
melody, c, 16thNote, mezzoForte,
polyphony, b, 8thNote, mezzoForte,
melody, f#, 16thNote, mezzoForte,
etc.
etc.
score (MusicXML)
extract values for each contextual factor and add them to deviational factors for each note as a label to build context-dependent models

Figure 2. Extraction of contextual factors

2.1 Context-dependent model

If the performance seems to be personalized, it is considered that the resultant personality is caused by biases in the trends of performance. The trend of deviation is observed as a distribution with some focus, according to deviations for each note extracted from each note o observed from the measured performance and the corresponding score (see Figure 1). We can think of the model as a Gaussian probability density function (PDF) so as to approximate the behavior of deviations; this model is able to cope with complex behaviors according to the Gaussian mixture model (GMM) approach. The PDF $\mathcal{N}$ of the observation vector $\boldsymbol{o}$ is defined by

$$\mathcal{N}(\boldsymbol{o}_m|\boldsymbol{\mu}_m,\boldsymbol{\sigma}_m) = \frac{1}{\sqrt{(2\pi)^D \Pi_{d=1}^{D} |\sigma_{md}|}} \exp\left(-\frac{1}{2}\sum_{d=1}^{D}\frac{(o_d-\mu_{md})^2}{\sigma_{md}}\right), \tag{1}$$

where $\boldsymbol{o}$ is observed with D deviational factors, o_d is the dth dimension for observation vector $\boldsymbol{o}$, m is the mixture index of the M Gaussian component densities, $\boldsymbol{\mu}$ is the mean vector, and $\boldsymbol{\sigma}$ is the diagonal covariance matrix.

However, the cause of the deviating behavior is not considered in this model. The performance of musical instruments consists of playing the sequences of notes according to the score. Therefore, it is obvious that the qualities of each note have some musical significance. As a general example, we consider performing two notes with different representations in terms of dynamics. In this case, the amount of deviation between them may be differ not only in the dynamics, but also in the timing, because of their expressive representations. Also, the extent to which the performer deviates from the average for the note with the representation is considered to be under the influence of some individuality. In the past, there were several studies that attempted to estimate the performers' characteristics by referring to the amount of deviation in timing and dynamics [5–7]. However, it is also necessary to consider what kind of representation leads to such behavior, using some musical knowledge that supersedes the mixture in the GMM.

Several factors complicate the process of occurrence. We make the following considerations to organize this subject:

- The performer obtains information from the musical score, and then creates his/her own interpretation using that information, thus introducing deviations into the performance.
- The trend of deviations occurring is also influenced by unintentional factors such as the performer's physical limitations.

We believe that the latter factor is not necessary, because it is considered likely based on relatively simple arguments, and the progress of performance technology is a means to reduce the interference of factors, such as unintentional representations. Additionally, factors (such as the former) influence the occurrence of this deviation, which is considered significant because it is intended to expand the range of expression in accordance with technological progress. However, criteria tend to be abstract and difficult to qualify, even for the performers themselves. Therefore, we do not directly address the interpretation of the music itself. Instead, we associate the trends in the deviation with the expressive representations, which affects the performer's musical interpretation.

All the information used here is in the form of unambiguous values that are available in the score, such as pitch, note value, dynamics, and so on, because we want to eliminate any undefined properties throughout the process. There is also the musical phrase to consider, which has some relationship that holds among surrounding notes. We introduce them under the term "context." Models in which context is applied are called "context-dependent," because they construct a kind of context that contributes to the interpretation. The parameters of the model are the same as the model mentioned above; however, each model has its own combination of contexts that is dealt with individually (see Figure 2). The description of the behavior for each model can be simplified because it is defined by a number of combinations. Therefore, each model is trained using a single Gaussian component density, as shown in Equation (1) .

2.2 Tree-based clustering

The purpose of introducing context is to associate a performer's interpretation of the musical composition with the deviations in the performance. A more detailed representation of the information obtained from the score has to con-

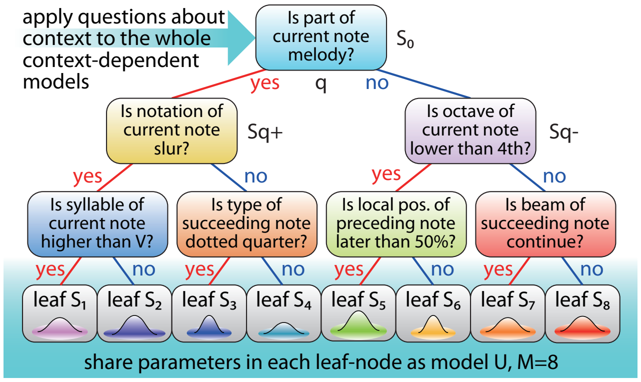

apply questions about context to the whole context-dependent models
Is part of current note melody?
S₀
yes
q
no
Is notation of current note slur?
Sq+
Is octave of current note lower than 4th?
Sq-
yes
no
yes
no
Is syllable of current note higher than V?
Is type of succeeding note dotted quarter?
Is local pos. of preceding note later than 50%?
Is beam of succeeding note continue?
yes
no
yes
no
yes
no
yes
no
leaf S₁
leaf S₂
leaf S₃
leaf S₄
leaf S₅
leaf S₆
leaf S₇
leaf S₈
share parameters in each leaf-node as model U, M=8

Figure 3. Example of a decision tree

sider a variety of contexts. However, with increasing use of contexts, the quantity of combinations of contexts increases exponentially. This effect is detrimental to model training, because the training data for each model will be significantly reduced. On the other hand, fragmented information has little meaning by itself. Therefore, it is necessary to classify a large number of combinations of contexts at a scale that matches the performer's significant interpretation. However, it is beyond human power to decide appropriate criteria for each case of classification. To address these issues, a method is necessary to reconstruct and decompose models efficiently, and to capture the varied expressive representations obtained from the score. We use tree-based clustering [4] to classify the context-dependent models.

Tree-based clustering divides all possible combinations of context-dependent model into a countable number of clusters. As a result, a decision tree (a binary tree in which a question is attached to each node) is obtained. In this method, each of the questions relates to the contextual factors for the preceding, current, and succeeding note. One tree is constructed for each deviational factor so as to cluster all of the corresponding behaviors of all context-dependent models. This is done because there are different trends of behavior for each deviational factor. All context-dependent models in the decision tree are divided into M nodes by clusters $S_1, \cdots, S_M$, such that one model $U(S_1, \cdots, S_M)$ is defined for each leaf node. For example, the tree shown in Figure 3 will partition its behaviors into eight subsets with the same number of leaf nodes. The questions and topology of the tree are chosen so as to maximize the likelihood of the training data, given these tied behaviors, by estimating the parameters of a Gaussian PDF. Once these trees have been constructed, data with unseen contexts can be classified in any leaf node by tracing the questions in the tree.

Initially, all the context-dependent models to be clustered are placed at the root node of the tree. The log likelihood of the training data is calculated, supposing that all of the models in that node are tied. Then, this node is divided into two by finding a question that divides the model in the parent node such that the log likelihood (maximally) increases. The log likelihood L for node S_m is given by

$$L(S_m) = -\frac{1}{2}\Gamma_m(K + K\log(2\pi)L\log|\Sigma_m|), \quad (2)$$

where Γ_m is the amount of data for training at node S_m. This process is then repeated by dividing the node in a way that creates the maximum increase of log likelihood until the minimum description length (MDL) criterion [9] is met. This step is carried out to optimize the number of clusters without using external control parameters. In order to optimize the size of the tree, we use an algorithm with a pragmatic cost of computation. Here, let us assume that node S_m of model U divides into two nodes, S_{mq+} and S_{mq-}, by answering question q. Then, let $\Delta_m(q)$ be the difference between the description length after division and before division, that is $l(U') - l(U)$. The description length of model U' is represented by the following equation:

$$\begin{aligned} I(U') = &\sum_{m'=1,\neq m}^{M} \frac{1}{2}\Gamma_{m'}\left(K + K\log(2\pi) + \log|\Sigma_{m'}|\right) \\ &+ \frac{1}{2}\Gamma_{mq+}\left(K + K\log(2\pi) + \log|\Sigma_{mq+}|\right) \\ &+ \frac{1}{2}\Gamma_{mq-}\left(K + K\log(2\pi) + \log|\Sigma_{mq-}|\right) \\ &+ K\left(M+1\right)\log W + C, \quad (3) \end{aligned}$$

where $W = \sum_{m=1}^{M}\Gamma_m$, and C is the length of code required to choose a model (assumed here to be a constant value). The number of nodes in U' is $M+1$, Γ_{mq+} is the occupancy count for node S_{mq+}, and Γ_{mq-} is that of node S_{mq-}. The difference $\Delta_m(q)$ is given by

$$\begin{aligned} \Delta_m(q) &= l(U') - l(U) \\ &= \frac{1}{2}(\Gamma_{mq+}\log|\Sigma_{mq+}| + \Gamma_{mq-}\log|\Sigma_{mq-}| \\ &\qquad - \Gamma_m\log|\Sigma_m|) + K\log\sum_{m=1}^{M}\Gamma_m. \quad (4) \end{aligned}$$

When dividing models, we first determine the question q' that minimizes $\Delta_{0q'}$ and that is used at root node S_0. If $\Delta_0(q') < 0$, node S_0 is divided into two nodes, S_{q+} and S_{q-}, and the same procedure is repeated for each of these two nodes. This process of dividing nodes is carried out until there are no nodes remaining to be divided. If $\Delta_0(q') > 0$, then no dividing is executed.

3. EXPERIMENTS

In this section, we apply the method mentioned above to the real-measured performance data to verify its efficacy of using expressive representations from the musical score as priori information. This information is applied to the issue of classifying the trends of the deviational behavior during the musical performance.

3.1 Data of real-measured expressive performance

Experiments in this paper use expressive performance data from a database ([1] and original data we collected). These contain information of musical expression on experts' expressive piano solo performances of classical Western musical compositions. The data of performance used in the experiments are as follows:

- performers

  PA V. D. Ashkenazy
  PG G. H. Gould
  PP M. J. Pires
  PR S. T. Richter
  PX Five anonymous semi-professional performers

- referred scores

  SBI J. S. Bach: "Two part Inventions BWV 772–786," Henle Verlag, pp. 2–31.
  SBW J. S. Bach: "The Well-Tempered Clavier BWV 846," Wiener Urtext Edition, pp. 2–3.
  SCN F. F. Chopin: "Nocturne No. 10," Paderewski Edition, pp. 54–55.
  SM3 W. A. Mozart: "Sonata K. 331, the First movement," Wiener Urtext Edition, pp. 18–27.
  SM5 W. A. Mozart: "Sonata K. 545, the First movement," Henle Verlag, pp. 266–269.

The actual performances also include notes do not correspond to the score. The current form of our method excludes these notes from the data used to train the model.

3.2 Design of models

The values of deviations and contexts are extracted by comparing the performance and the score, as shown in Figure 1 and Figure 2. The five factors in which there could be deviation (shown below) are extracted for each note; therefore, the dimensionality $D = 5$ in Equation (1).

- Factors that depend on the note:

  onset Timing when striking the key. The amount of deviation is represented relative to a beat. If the performed note is struck one half beat faster, the deviation of onset is -0.5.
  offset Timing when releasing the key, represented in the same way as the deviation of onset.
  gate time The quotient of the time taken to depress the key in the performance divided by its length on the score. If both are exactly the same, the deviation of gate time is 1.
  dynamics Strength when striking the key, obtained in the same way as the deviation of gate time.

- Factor that depends on the beat:

  tempo Temporal change of BPM (current beat/average).

The contextual factors attached to context-dependent model are shown below. They are used for question to construct decision trees. In this experiment, the total number of questions used amounted to more than two thousands.

- Extracted for $\{$preceding, current, succeeding$\}$ notes:

  syllable Interval name of the note and the tonic, i.e., minor third, perfect fifth, etc.
  step One of the twelve note names, from C to B.
  accidental Existence and type of accidental.
  octave Rough pitch of the note.
  chord Whether the note belongs to any chord.
  type Note value of the note.
  staff Clef and stage on the great staff the note is written in.
  beam Type of the note's beams, i.e., begin, continue, end, etc.
  local The note's position on the beat in the bar, represented as a percentage.

- Extracted for current note only:

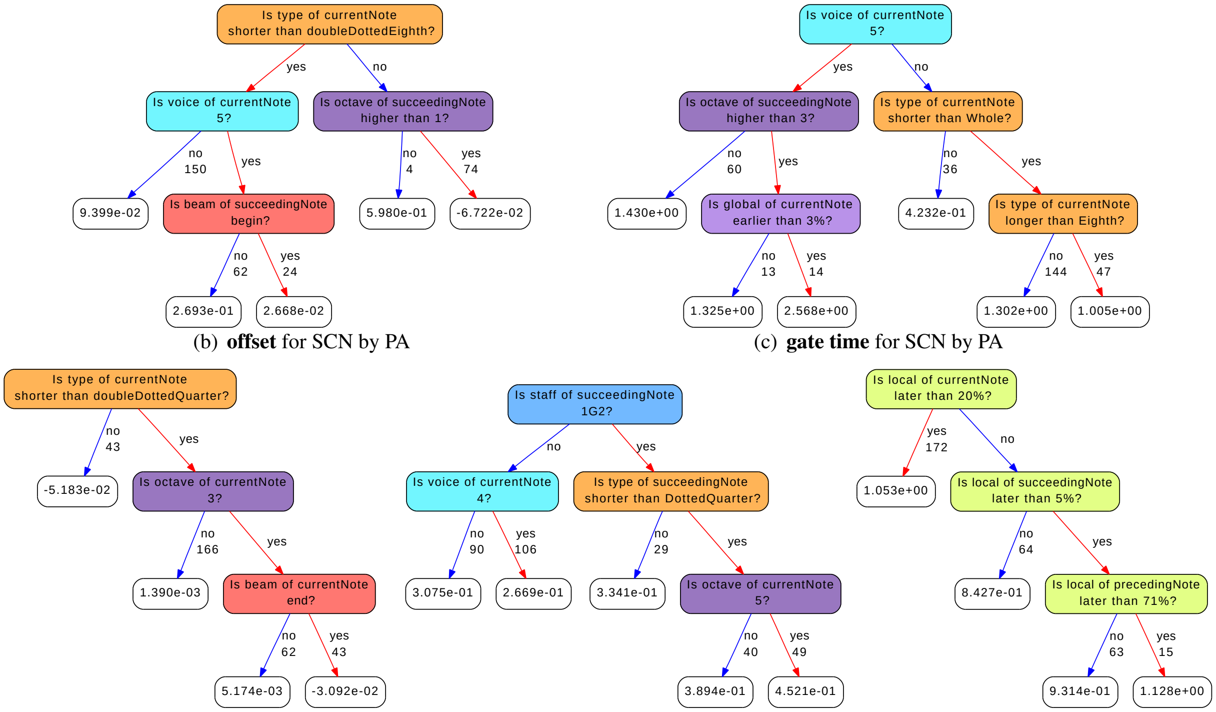
  global The note's position in elapsed time in the musical composition, represented as a percentage.
  voice Voice part of the note, defined by the author of the database.
  notations Noted signs for the note, such as dynamics, intonation, etc.

3.3 Efficacy of tree-based clustering

The tree-based clustering itself is an existing method; however, the effect of applying this method to a musical performance is unknown. Therefore, it is necessary to determine whether changes in generative efficiency can be seen in the bottom-up clustered model without additional information. To achieve concrete results, we tried to identify the performer from the performance data using the models. The data sets used in this case were SBI and SM3, both of which were performed by PX. The models were trained with the data of the compositions, which amounted to approximately one quarter of the data set. The tests used each datum of the remaining compositions in the same set; the percentage of the right choices for the performer by the trained model was calculated (called the rate of identification). Evaluation of resistance to the unseen data was also carried out using this test, as all models were tested with data that is not used to train the models. We differentiate these methods:

Tree-based clustering The model using the proposed method.
Bottom-up clustering The model trained by GMM with the same number of mixtures M as the leaves in the trees generated by tree-based clustering, and using the same data set that is used to train the models.

The result is shown in Figure 4, and the ratio of accuracy to the average of 20 ordinary human listeners for each method is also indicated in parentheses. This is a severe condition, and the most human listeners cannot tell the difference. However, proposed method can determine such subtle difference with high precision, because the ratio of *Tree-based* is about 232% for human listeners. Furthermore, the ratio of *Tree-based* for *Bottom-up* is about 111%. Therefore, it is confirmed that the accuracy can be improved upon to generate models that can respond to unseen data by using the clustering with the information from the score.


Is type of currentNote shorter than doubleDottedEighth?
yes
no
Is voice of currentNote 5?
Is octave of succeedingNote higher than 1?
no 150
yes
no 4
yes 74
9.399e-02
Is beam of succeedingNote begin?
5.980e-01
-6.722e-02
no 62
yes 24
2.693e-01
2.668e-02
Is voice of currentNote 5?
yes
no
Is octave of succeedingNote higher than 3?
Is type of currentNote shorter than Whole?
no 60
yes
no 36
yes
1.430e+00
Is global of currentNote earlier than 3%?
4.232e-01
Is type of currentNote longer than Eighth?
no 13
yes 14
no 144
yes 47
1.325e+00
2.568e+00
1.302e+00
1.005e+00
Is type of currentNote shorter than doubleDottedQuarter?
no 43
yes
-5.183e-02
Is octave of currentNote 3?
no 166
yes
1.390e-03
Is beam of currentNote end?
no 62
yes 43
5.174e-03
-3.092e-02
Is staff of succeedingNote 1G2?
no
yes
Is voice of currentNote 4?
Is type of succeedingNote shorter than DottedQuarter?
no 90
yes 106
no 29
yes
3.075e-01
2.669e-01
3.341e-01
Is octave of currentNote 5?
no 40
yes 49
3.894e-01
4.521e-01
Is local of currentNote later than 20%?
yes 172
no
1.053e+00
Is local of succeedingNote later than 5%?
no 64
yes
8.427e-01
Is local of precedingNote later than 71%?
no 63
yes 15
9.314e-01
1.128e+00

(b) **offset** for SCN by PA

(c) **gate time** for SCN by PA

(a) **onset** for SCN by PA

(d) **dynamics** for SCN by PA

(e) **tempo** for SCN by PA

Figure 5. Examples of structural and statistical differences in tree-structures for each deviational factor

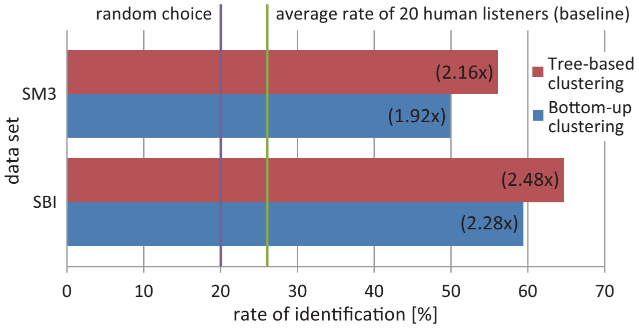


Figure 4. Results of identification test

3.4 Observation of decision trees

Next, we observe the decision trees obtained from the performance data to verify the kind of questions that divide the models and the statistical attributes of each model. The set of training data used here was SCN, performed by PA. Examples of the portion of the trees near the root are shown in Figure 5. Each node has the content of the question, each leaf gives the average deviation, and the number of models involved in each leaf is indicated by an arrow.

The trees of deviational factors belong to the timing (*onset*, *offset*, and *gate time*) have affinities in the kind of questions. The tree of dynamics also has the sequence of questions with the same contexts as the factors mentioned above; however, the kind of question on the root node is not seen. Although they have certain unique points, they have a similar structure. On the other hand, the tree of tempo has very different trends, both in terms of structure and questions.

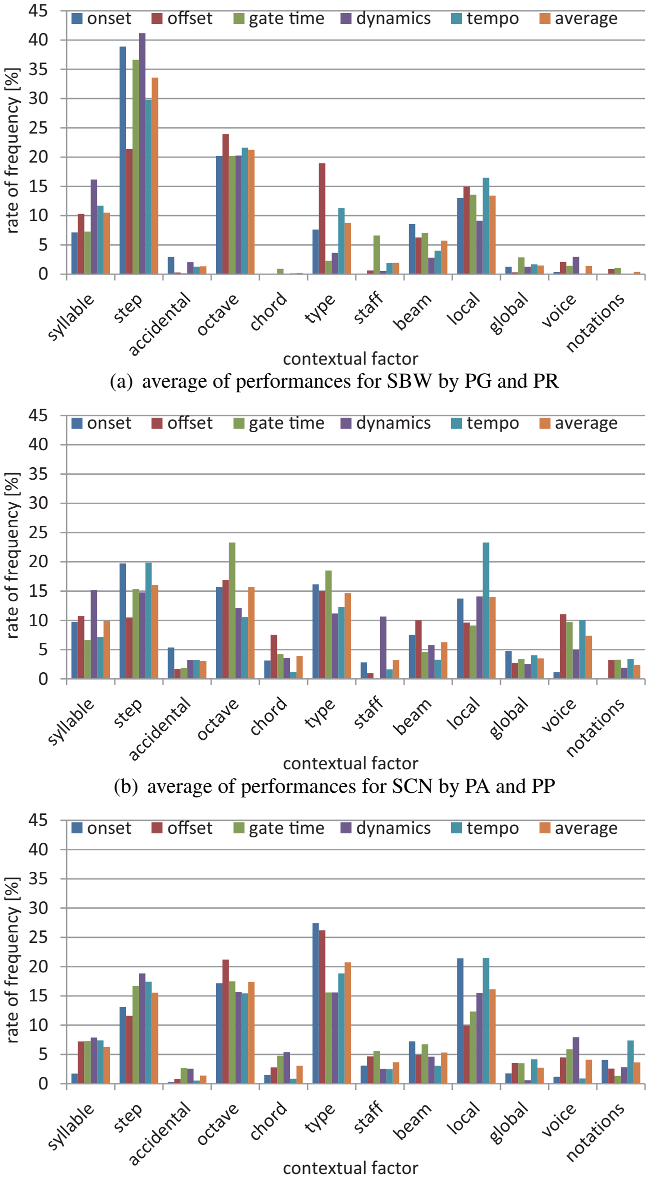
3.5 Contribution of contextual factors to decision trees

Due to the limitations of the available data, a more efficient analysis is needed to understand the trends of these factors. We therefore investigated the frequency of any question to find the degree of contribution to the trend of deviation caused by each contextual factor. The contribution C for contextual factor Q in a tree with M leaf nodes is counted by

$$C_Q = \sum_{m=1}^{M} \left(\frac{N_m}{N_{all}} \times R_Q \right), \tag{5}$$

where N_m is the number of context-dependent models shared by the mth leaf node, and R is the number of nodes related to Q in the path from the root node to the mth leaf node. The training data used here was SBW-by-{PG, and PR}, SCN-by-{PA, and PP}, and SM5-by-{PG, and PP}. The results for each composition are shown in Figure 6; we propose that these results show the priorities of performers' criterion to differentiate the behavior in the performance.

The trend of contextual factors that make a large contribution is the same in all compositions (e.g., *step*, *octave*, *type*, *local*, and *syllable*). We consider the essential part of the trees' construction to depend upon the selection order of these factors. On the other hand, the difference between offset and gate time is small, as mentioned above; however, these result shows some differences (for example, they are found in *step*, *octave*, and *type*). There is a possibility to reveal the diverging points of the deviations with expressive representations by observing more detailed classifications.



(a) average of performances for SBW by PG and PR

(b) average of performances for SCN by PA and PP

(c) average of performances for SM5 by PG and PP

Figure 6. Frequencies of contextual factors for each composition

4. CONCLUSIONS

In this paper, we presented a method for describing the characteristics of human musical performance. The experimental results of performer identification showed the use of the expressive representations from the musical score enables the efficient acquisition of the model of the performance. The results also showed that the proposed model can capture the characteristics of the performance from any subtle differences that cannot be found by most human listeners. Therefore, the efficacy of using expressive representations from the musical score to describe the characteristics of the musical performance was shown. This method can automatically learn the knowledge necessary to describe the tree structure of the model directly from the data of the performance. We believe that the availability of such objective elements from the proposed model is effective for the analysis of the performance. In the future, we will make comparisons based on more common and more extensive examples, in addition to attempting to improve the modeling method. Furthermore, this method can be applied to generate unseen performances. We are also making efforts in that direction.

5. ACKNOWLEDGEMENT

This research was partially supported by NIT president's discretionary expense for young researchers and a Grant-in-Aid for Young Scientists (B) from the Ministry of Education, Culture, Sports, Science, and Technology, Japan.

6. REFERENCES

[1] M. Hashida, T. Matsui, and H. Katayose: ”A New Music Database Describing Deviation Information of Performance Expressions,” *Proceedings of the International Symposium on Music Information Retrieval*, pp. 489–494, 2008.

[2] A. Kirke and E. R. Miranda: ”Survey of Computer Systems for Expressive Music Performance,” *Journal of ACM Computing Surveys*, Vol. 42, No. 1, Article 3, 2009.

[3] J. Langner and W. Goebl: ”Visualizing expressive performance in tempo-loudness space,” *Computer Music Journal*, Vol. 27, No. 4, pp. 69–83, 2003.

[4] J. J. Odell: ”The Use of Context in Large Vocabulary Speech Recognition,” Ph.D thesis, Cambridge University, 1995.

[5] B. H. Repp: ”A microcosm of musical expression: I. Quantitative analysis of pianists' timing in the initial measures of Chopin's Etude in E major,” *Journal of the Acoustical Society of America*, Vol. 104, No. 2, pp. 1085–1100, 1998.

[6] B. H. Repp: ”A microcosm of musical expression: II. Quantitative analysis of pianists' dynamics in the initial measures of Chopin's Etude in E major,” *Journal of the Acoustical Society of America*, Vol. 105, No. 3, pp. 1972–1988, 1999.

[7] B. H. Repp: ”A microcosm of musical expression: III. Contributions of timing and dynamics to the aesthetic impression of pianists' performances of the initial measures of Chopin's Etude in E major,” *Journal of the Acoustical Society of America*, Vol. 106, No. 1, pp. 469–478, 1999.

[8] C. S. Sapp: ”Comparative analysis of multiple musical performances,” *Proceedings of the International Symposium on Music Information Retrieval*, pp. 497–500, 2007.

[9] K. Shinoda and T. Watanabe: ”MDL-Based context-dependent subword modeling for speech recognition,” *A. Acoustical Society Japan (E)*, Vol. 21, No. 1, pp. 70–86, 2000.

[10] G. Widmer: ”Machine discoveries: A few simple, robust local expression principles,” *Journal of New Music Research*, Vol. 31, No. 1, pp. 37–50, 2002.

[11] G. Widmer, S. Dixon, W. Goebl, E. Pampalk, and A. Tobudic: ”In search of the Horowitz factor,” *AI Magazine*, Vol. 24, No. 3, pp. 110-130, 2003.