MID-FiLD: MIDI Dataset for Fine-Level Dynamics

Jesung Ryu[1], Seungyeon Rhyu[1], Hong-Gyu Yoon[1], Eunchong Kim[1], Ju Young Yang[2], Taehyun Kim[1*]

[1]Pozalabs, Republic of Korea
[2]Duke University, United States
{jesung, seungyeon, honggyu, eunchong, taehyun}@pozalabs.com, juyoung.yang@duke.edu

Abstract

One of the challenges in generating human-like music is articulating musical expressions such as dynamics, phrasing, and timbre, which are difficult for computational models to mimic. Previous efforts to tackle this problem have been insufficient due to a fundamental lack of data containing information about musical expressions. In this paper, we introduce MID-FiLD, a MIDI dataset for learning fine-level dynamics control. Notable properties of MID-FiLD are as follows: (1) All 4,422 MIDI samples are constructed by professional music writers with a strong understanding of composition and musical expression. (2) Each MIDI sample contains four different musical metadata and control change #1 (CC#1) value. We verify that our metadata is a key factor in MID-FiLD, exerting a substantial influence over produced CC#1 values. In addition, we demonstrate the applicability of MID-FiLD to deep learning models by suggesting a token-based encoding methodology and reveal the potential for generating controllable, human-like musical expressions.

Introduction

Expressive dynamics is an essential element for improving the quality and completeness of music (Todd 1992). As shown in the spectrogram in Figure 1, expressive dynamics control the loudness of an instrument, which is a subjective perception of sound pressure. In a performance, the performer focuses not only on performing the correct pitch and duration of the notes but also on conveying musical expression including loudness intended by the composer (Jędrzejewska, Zjawiński, and Stasiak 2018). In order to create high-quality and realistic music using deep learning techniques, recent studies have aimed to mimic characteristics of human performances including expressive dynamics (Huang et al. 2018; Huang and Yang 2020). Furthermore, expressive dynamics has been regarded as a crucial factor in controllable generation of music (Tan, Luo, and Herremans 2020; Wu et al. 2022).

Duly recognizing the significance of conveying expressive dynamics, a majority of studies dealing with MIDI data focused on generating note-level MIDI velocity, which can adjust the loudness of each note (Cancino-Chacón et al. 2018; Jeong et al. 2019b; Lauly 2010). Likewise, in the area of audio synthesis, previous studies have made efforts to reflect the loudness of music when rendering MIDI data (Castellon, Donahue, and Liang 2020; Wu et al. 2022). To accomplish this, most studies extracted the loudness features directly from the frequency and amplitude of the audio signal without using domain-specific annotations.

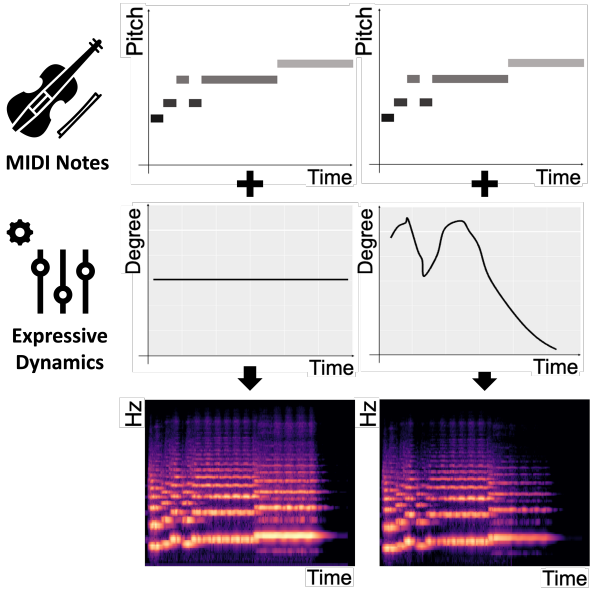

Figure 1: Example illustrating variations in Mel-spectrogram according to alterations in expressive dynamics. As the value of expressive dynamics fluctuates, the intensity of the sound signal varies over time. It can be seen that expressive dynamics affect the amplitude of a specific frequency in music, which leads to changes in loudness.

However, these previous attempts are bound by two inherent limitations. Primarily, note-level expressive dynamics are insufficient to reenact musical expression performed by most western instruments. Except for piano or percussion, in which the attack of a note defines the loudness, most western instruments can induce variation in the loudness of a single note. For example, a string, woodwind, or brass instrument can perform a crescendo that stretches within a note (Berndt

*Corresponding author.

Attributes	Instrument	Mood	Track role	Note-level Dynamics	Fine-level Dynamics
MAESTROv3 (2019)	Single	×	×	◯	×
MusicNet (2017)	Multiple	×	×	◯	×
URMP (2018)	Multiple	×	×	×	×
Lakh MIDI (2016)	Multiple	×	×	◯	△
ComMU (2022)	Multiple	×	◯	◯	×
MID-FiLD (Ours)	Multiple	◯	◯	◯	◯

Table 1: Comparison of MID-FiLD to other recent MIDI datasets. We compare MID-FiLD to other MIDI datasets on: instruments, metadata (mood and track role), and parameters of dynamics (note-level and fine-level). Since only a portion of Lakh MIDI dataset contains fine-level dynamics, it is marked as △.

and Hähnel 2010). Therefore, modelling fine-level expressive dynamics is necessary to create music with a variety of instruments. In addition, expressive dynamics can be produced based not only on the properties of the notes but also on the auxiliary attributes, or metadata, of music. Further, human composers tend to factor in various elements, including the type of instrument, mood, or track role of the music, when crafting expressive dynamics (Gabrielsson and Juslin 1996; Li et al. 2018). Hence, it can be a naive approach to simply extract musical dynamics from an audio signal of human performance without any metadata aligned. Although some MIDI datasets may include partial samples with annotations of fine-level dynamics (Raffel 2016), they are also limited in providing a sufficient amount of aligned metadata that elucidates the characteristics of the music.

In this paper, we introduce MID-FiLD, a new dataset containing 4,422 MIDI music samples that are paired with fine-level expressive dynamics and meta information collected by professional composers. We focus on fine-level annotations for expressive dynamics represented as parameter values of *modulation wheel* of MIDI, a type of control change messages within MIDI. Control change (CC) messages denote parameter values that modify the attributes of various instrumental sounds over time (Moog 1986). MID-FiLD provides #1 parameter values of CC (i.e., CC#1), corresponding to the incremental amount of modulation wheel to represent the time-varying dynamics of the instrument's sound (for details, refer to Appendix A). The CC#1 values for each MIDI sample were carefully annotated by domain professionals, which leads to annotations that are substantially more sophisticated than the note-level annotations or features from the existing datasets. Furthermore, in line with the preceding study (Lee et al. 2022), our dataset incorporates metadata including mood, track role, and min-max range of CC#1 values. Given the importance of their role in understanding musical conditions, we show that these metadata can be employed strategically in the dynamics generation task.

Based on our dataset, we conduct the following analysis and experiments to demonstrate its excellence. First, we analyze the dataset through exploratory data analysis (EDA) to reveal the relationships between expressive dynamics and metadata of music. Given that our dataset contains not only dynamics but also meta information which represents properties of music, it becomes feasible to reveal the tendency of dynamics that are otherwise unexplored within the confines of MIDI-only data. Second, we evaluate baseline models on a generation task with MID-FiLD through an ablation study with our novel, task-specific representation. By training a deep learning model with MID-FiLD, improved performances are observed in terms of fidelity and controllability.

In summary, the main contributions of our work are as follows:

- MID-FiLD contains fine-level expressive dynamics. The dataset is crafted by professional composers, which encompasses a broader range of natural loudness variations compared to the existing note-level dataset.
- This is the first dataset which contains both fine-level dynamics and sufficient metadata including track role and mood. Moreover, we identify a substantial correlation between metadata and dynamics through experiments, implying the strength of our dataset in tasks involving the generation of dynamics.
- We provide a baseline model including data representation, and corresponding metrics for generating fine-level values of dynamics with MID-FiLD.

Additionally, the main differences of MID-FiLD compared to existing datasets are summarized in Table 1.

Related Work

Datasets Containing Expressive Dynamics

A number of music datasets have provided information related to expressive dynamics through various attributes. Symbolic music datasets such as Lakh MIDI Dataset (Raffel 2016), MAESTRO (Hawthorne et al. 2019), Pop1k7 (Hsiao et al. 2021), or ComMU (Lee et al. 2022) include human performance data in the format of MIDI. With the exception of Lakh MIDI Dataset, these datasets only include note velocities as parameters of expressive dynamics, which are limited to note-level variation in loudness. Lakh MIDI Dataset comprises multi-track MIDI with various types of control changes for flexible control of distinctive instruments. However, it does not contain sufficient meta information such as track roles or mood-related attributes for each sample. On the other hand, audio datasets can provide high-quality audio

recordings for various instruments. RWC Music database (Goto et al. 2002), MusicNet (Thickstun, Harchaoui, and Kakade 2017), and URMP (Li et al. 2018) are examples of audio datasets collected through recording. Although these datasets can provide features related to dynamics from the raw audio, the quality of features depends on extraction methods that are either heuristic or automatic. These extracted features can be less accurate than those of careful annotations produced by professional musicians.

Modeling Expressive Dynamics of Symbolic Music

In the symbolic domain, expressive dynamics has been considered as one of the effective parameters to modify quantized MIDI events into realistic music performances (Cancino-Chacón et al. 2018). Finding good expressive dynamics has been often associated with generating relevant note velocity. Recent deep learning techniques, such as recurrent neural network (RNN) (Lauly 2010) and graph neural network (GNN) (Jeong et al. 2019b), encouraged generative models to learn non-linearity in a large number of expressive parameters conditioned by the musical score attributes including note pitches, durations, or onset timings. Moreover, deep probabilistic models such as variational autoencoder (VAE) have facilitated the stochastic generation of the expressive dynamics constrained by the musical score (Maezawa 2018; Maezawa, Yamamoto, and Fujishima 2019; Jeong et al. 2019a). Nonetheless, these studies are mostly limited to piano performance and note-level dynamics.

Modeling Expressive Dynamics of Audio

Expressive dynamics have been regarded as one of the intermediate parameters for synthesizing audio of human-performed music from MIDI data. While conventional approaches include a feed-forward neural network predicting the dynamics of a motif performed by a violin (Ortega, Perez-Carrillo, and Ramírez 2019), recent studies focused on Differentiable Digital Signal Processing (DDSP) (Engel et al. 2020). Utilizing LSTM-based model (Jonason 2020) or MIDI2Params model (Castellon, Donahue, and Liang 2020), high-level acoustic parameters including loudness from a given MIDI were predicted as inputs for the DDSP module. More recently, MIDI-DDSP has become state-of-the-art in realistic audio synthesis for various instruments from MIDI scores (Wu et al. 2022). MIDI-DDSP exploits an expression generator, which aims to predict 6 hand-crafted parameters related to volume, pitch, and noise from a MIDI input. Although it allows a user to finely adjust time-varying sound attributes, its application is limited to note-level dynamics. Furthermore, achieving accurate and sensitive integration of the parameters necessitates a deep understanding of music and audio signals on the user's part.

MID-FiLD

Data Collection

MID-FiLD has 4,422 samples that consist of short note sequences with 4 corresponding metadata. Table 2 includes basic information of MID-FiLD.

# Samples	4,422
# Notes	91,863
# Average notes per sample	20.8
Types of instruments	18
Types of track role	6
Types of mood	19
Range of CC#1 value	0-127

Table 2: Basic information of MID-FiLD.

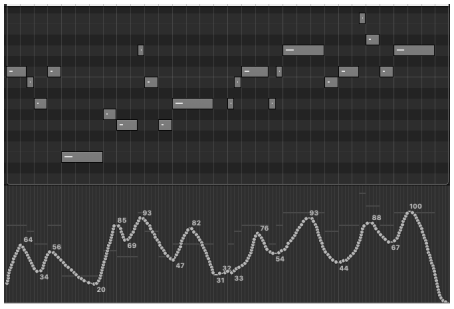

Figure 2: Example of CC#1 values with notes in MID-FiLD displayed by a digital audio workstation (Logic Pro X). Ascending values express an increase in loudness, and descending values express a decrease in loudness.

MID-FiLD was made by professional composers based on a systematic approach. Composers were divided into two groups. One group created a composition guideline for the MIDI samples including metadata of each sample, and the other composed MIDI samples based on the guideline, following the data collection workflow in ComMU dataset (Lee et al. 2022).

In addition to creating MIDI notes and entering metadata, composers also incorporated nuanced dynamics into each sample, taking into account not only the melody but also the sample's metadata. Effectively applying their sense and knowledge in musical composition, composers drew semi-continuous values of expressive dynamics using MIDI control change messages. We selected control number 1, among other control change messages, to explicitly represent expressive dynamics in a consistent manner. The control number of the control change message corresponds to information related to performance controls such as wheels or pedals: our dataset utilized control number 1 (CC#1), modulation wheel. Figure 2 shows a graph of CC#1 value with notes in a sample over time, captured as a screenshot within a digital audio workstation. The range of CC#1 value spans from 0 to 127, which fluctuates as time-series data in alignment with melody.

Metadata

In this section, we take a closer look at the definition of the 4 metadata.

Instrument. Our taxonomy of instruments follows that of Western musical instruments (Kartomi 1990). Based on the

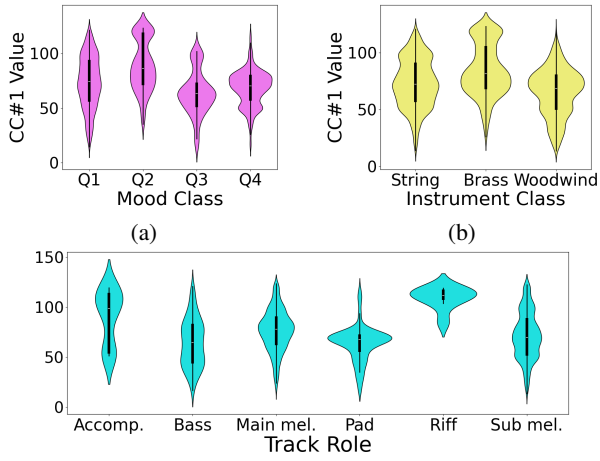

Figure 3: Data distributions of CC#1 values by each metadata group of (a) mood, (b) instrument, and (c) track role.

general classification system of Western instruments, MID-FiLD has 18 different instruments belonging to either bowed string, woodwind, or brass. Such distribution of instruments has a meaningful effect on fine-level control of dynamics because the strength of bowing and blowing can increase or decrease regardless of the initial velocity of a note, which corresponds to mere note-level control.

Track role. With reference to ComMU dataset (Lee et al. 2022), track role is a classification of note sequences, based on its role in a piece of multi-track music. We divide multi-track into main melody, sub-melody, accompaniment, bass, pad, and riff, where each group has different note figures containing its own characteristics of the track.

Mood. Mood is a sentiment which represents the atmosphere of a sample. Based on information about the audio key (major, minor), tempo, traits of rhythm, and other descriptive guidelines, composers standardize the taxonomy of different moods based on similar keywords for the atmosphere of each sample. As a result, the mood is classified into 19 groups, characterizing the melody's emotion in each sample.

Min&max CC#1 value. In order to provide additional information regarding the highest/lowest value of CC#1 while training, we extract minimum and maximum CC#1 values as metadata. Since we tokenize the information of metadata, the CC#1 value range (0-127) is quantized into a discrete range with a 5-bin size, allowing us to reduce its complexity and benefit from the regularization effect when preprocessing.

Data Analysis

In this section, we further investigate MID-FiLD with respect to metadata. In particular, we examine relationships between CC#1 value and metadata (i.e. mood, instrument, and track role) in terms of data distribution. Distinguishable

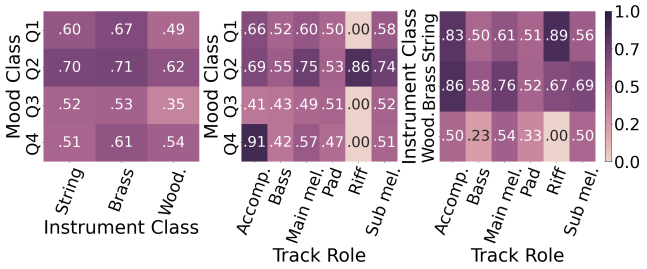

Figure 4: Heatmaps illustrating relationships between pairs of metadata. Each entry denotes the intersection of the average CC#1 value of the two subsets from two different metadata.

data distribution according to metadata is important in generating diverse music (Lee et al. 2022). Hence, we verify whether there are observable differences in the data distribution of CC#1 value based on given metadata.

Our dataset contains sufficient labels of metadata including mood, instrument, and track role. To demonstrate clear data distribution by each metadata, we group labels of mood and instrument into higher-level classes. For mood, we refer to Russell's 4Q model (Russell 1980), which classifies emotion into one of the quadrants comprised of valence and arousal axes. To this end, we assign each of the 19 mood labels to one of the four valence-arousal classes according to previous studies for classifying music emotion (Levy and Sandler 2007; Bischoff et al. 2009; Laurier et al. 2009). Instruments are divided according to the Western instrument categories (string(bowed), woodwind, and brass instruments). Track roles are analyzed without grouping. To see details on the grouping of the labels, refer to Appendix B.2.

Distribution by Each Metadata

We average CC#1 values of each sample. To measure the significance of differences among data distributions, we conducted Welch's ANOVA and Games-Howell test as the post hoc pairwise comparison. Figure 3 illustrates the distributions of CC#1 mean values of different groups of metadata. Mood classes show significant differences in distribution from one another ($p < 0.01$), where Q2 shows the highest mean of CC#1 values compared to the other quadrants. This indicates that MIDI note samples at high arousal and low valence can induce increased dynamics, duplicating previous studies that negative arousal is deeply related to sound intensity (Gomez and Danuser 2007; Hung et al. 2021; Weninger et al. 2013). Instrument classes also show meaningful pairwise differences in CC#1 values ($p < 0.0001$), where the brass class has the highest mean of CC#1 values (Phillips and Mace 2008). Track role, however, shows significant pairwise differences excluding accompaniment vs. main melody pair ($p > 0.1$), accompaniment vs. riff pair ($p > 0.1$), and bass vs. pad pair ($p > 0.5$). Meanwhile, bass and pad share significantly analogous variances in a low range of CC#1 values due to their similar roles within multi-track music (Lee et al. 2022).

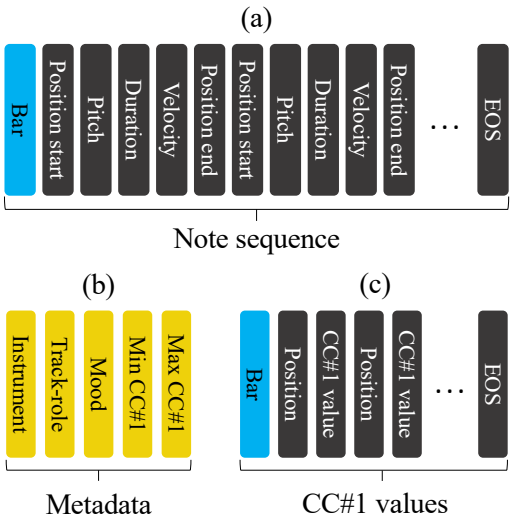

Figure 5: Illustration of our input encoding approach for MID-FiLD. (a) Notes are decomposed into the following five components: *position* (*start*, and *end* each), *pitch*, *duration*, and *velocity* between repeated bar tokens. (b) For metadata, each field out of five results in a single token representation. (c) For CC#1 values, each occurrence on MIDI is decomposed to *position* and *value* that alternates between repeated bar tokens.

Relationships Among Metadata

Figure 4 shows the aggregated mean of CC#1 values of each pair of metadata, normalized to the range of [0,1] for clarity of presentation.

The value "0" denotes the absence of data for the corresponding intersection of the two groups. First, Q2 has the highest mean regardless of the instrument, whereas brass has the highest mean of CC#1 among instruments regardless of mood. Woodwind displays the smallest CC#1 regardless of track role and mood. Meanwhile, bass and pad have decreased dynamics compared to other track roles independent of instruments and mood. Moreover, the intersection of accompaniment and Q4 shows the highest mean of CC#1, while that of accompaniment and Q3 shows the lowest. This signifies that the dynamics in accompaniment can largely vary based on mood rather than instrument. Generally, it is evident that mood, instrument, and track role interact with one another to some extent in relation to CC#1 value. At the same time, other explicit tendencies in CC#1 values in each metadata are observed, suggesting a controllability of fine-level dynamics through metadata.

Experimental Setup

In this section, we investigate the potential of our dataset by demonstrating an implementation task on a generative model through an objective evaluation method. The basic gist of our generative task is to predict a CC#1-time series value appropriate for the note sequence and corresponding metadata of a given sample. The methodologies of our experiments including self-defined evaluation metrics are described below in detail.

Implementation

We devised a token-based representation for the input encoding process of MID-FiLD, which is a general approach inspired by Natural Language Processing to handle symbolic music data. Note sequences and their associated metadata and CC#1 values are tokenized and encoded into a sequence of integers based on specific pre-defined mappings. For details on the input encoding vocabulary, refer to Appendix C.1.

Input encoding. Figure 5 shows our tokenization strategy, which inherits the methodology of REMI representation(Huang and Yang 2020). Note sequence representations are acquired by tokenizing note components into *position*, *pitch*, *velocity*, and *duration*. One notable difference between REMI and our representation is that a new *position end* token class has been added. Then the mathematical formulation of the encoded note sequence can be written as follows:

$$X = \{x_1^B, x_2^{P_S}, x_3^H, \ldots, x_{K-2}^D, x_{K-1}^V, x_K^{P_E}\} \quad (1)$$

where each B, P_S, H, D, V, and P_E represents *bar* (in music score), *position start*, *pitch*, *duration*, *velocity*, and *position end*. Note that K represents the number of tokens in the encoded note sequence. For meta information, each field is encoded to a single metadata token m_i. Thus simply:

$$M = \{m_1, m_2, \ldots, m_5\}. \quad (2)$$

Similarly, formulation of CC#1 values can be described as

$$C = \{c_1^B, c_2^P, c_3^V, \ldots, c_{N-1}^P, c_N^V\}, \quad (3)$$

where B, P and V indicate *bar*, *position* (of CC#1 value), and *value* respectively. N stands for the number of tokens in the CC#1 value sequence. The tokens for encoding position in CC#1 sequence share the same embedding space with those of position in note sequence while sharing the same resolution rate of 128 per measure. Such traits enable the model to learn fine-level control through high-resolution representation.

Problem definition. Our task focuses on generating proper CC#1 values conditioned by specific metadata and a given note sequence. We demonstrate our task by using the vanilla transformer model(Vaswani et al. 2017) with its encoder and associated auto-regressive decoder. Let (X, Y) be the input pair of encoder and decoder for our model. We construct our encoder input X with note sequence directly as in Eq. (1), and decoder input Y by concatenating the metadata token sequence (2) and the CC#1 token sequence (3), i.e.,

$$\begin{aligned} Y &= \mathrm{Concat}(M, C) \\ &= \{m_1, \ldots, m_5, c_1^B, c_2^P, \ldots, c_N^C\}. \end{aligned} \quad (4)$$

Now, to learn the sequence of CC#1, the decoder can be trained by minimizing the following negative auto-regressive log-likelihood of the sequence $Y := \{y_t\}_{t=1}^T$,

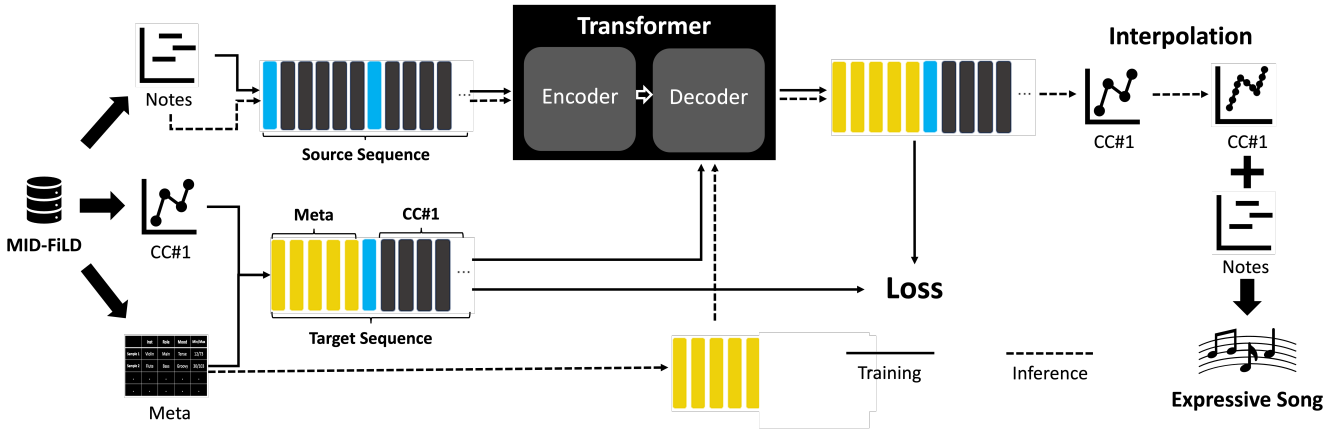

Figure 6: Overall process of the CC#1 generation task. During training, note information is used as the source of the model, and CC#1 values with metadata are used as the target of the model. The model updates its parameter using the loss (negative log-likelihood) between the target sequence (ground truth) and the generated sequence (predicted). By forcing the target sequence to have a specific initial metadata sequence, the model can generate the CC#1 values conditioned by the note (source) sequence and metadata.

conditioned by X and M:

$$\mathcal{L}_{\theta,\phi}(Y) = -\sum_{t=6}^{T} \log p_\theta(y_t \,|\, y_{<t},\, \mathcal{E}_\phi(X)), \qquad (5)$$

where $\mathcal{E}_\phi(\cdot)$ is the encoder. Note that the index starting from $t = 6$ in the summation assures the conditioning on metadata M. After the training phase, we can generate the target CC#1 sequence from given notes $\bar{X}$ and its metadata $\bar{m}_1, \ldots, \bar{m}_5$ by using the following sampling process

$$\widehat{y}_{t+1} \sim p_{\theta_*}(\,\cdot \,|\, \widehat{y}_{\leq t}, \mathcal{E}_{\phi_*}(\bar{X})), \quad t > 5, \qquad (6)$$

where $\widehat{y}_1, \ldots, \widehat{y}_5$ is forced as $\bar{m}_1, \ldots, \bar{m}_5$. While the sampling method can be arbitrarily chosen, we used Top-k sampling ($k = 32$).

Training & inference. Figure 6 illustrates the complete pipeline of our task including the training and inference phase. We divided 4,422 MID-FiLD samples into a ratio of 8:1:1 for training, validation and test sets respectively, leaving about 10% of the data for test performance measurements (i.e., $3,547 : 443 : 432$). The test set is randomly sampled in a stratified manner to follow the distribution of the training set with respect to its instruments. During the training phase, each tokenized input pair (X, Y) as in Eq. (1) and (4) is repeatedly delivered to the model, optimizing the loss (5) until saturation (un-dashed arrows). In the inference phase, the trained decoder p_{θ_*} generates the CC#1 values through conditional auto-regressive sampling as in Eq. (6), until the EOS token is returned (dashed arrows). The generated tokens are converted into a time series of CC#1 values by following the exact inverse process of the input encoding, and additional linear interpolation is applied to the generated series for adequate rendering into an expressive song.

Evaluation Metrics

Fidelity. We evaluate the difference between CC#1 values from the generated sample and those of the ground truth sample which is drawn by professional composers. Generating a realistic human-like sample requires a minimal gap between the generated value and the ground truth value. Generated CC#1 values of each sample can be treated as semi-continuous data, where the difference of value (at y-axis) at each time point (at x-axis) defines fidelity as **RMSE** score after interpolation. This assumes that generated CC#1 values are linearly interpolated before calculating the difference of values at each time point from the ground truth values. To remove the model's uncertainty considering the property of the metric, we utilize the greedy approach instead of top-k sampling when measuring the score in this metric.

Controllability. We define the controllability based on two facets as follows:

- **Differences of min. & max.:** We calculate the gap between the minimum (maximum) values in the generated sample and that in the ground truth sample, which can be defined as the difference of min. (max.). Such value indicates how close the model-generated value is to the intended min/max range. Let set of CC#1 values in the ground truth sample as $C_t = \{v_1, v_2, ..., v_n\}$, and set of CC#1 values in the model-generated sample as $C_g = \{v_1, v_2, ..., v_m\}$. Then, the difference of each can be defined as follows:

$$D_{\min} = |\min(C_t) - \min(C_g)| \qquad (7)$$

$$D_{\max} = |\max(C_t) - \max(C_g)| \qquad (8)$$

- **Mood classification accuracy (4Q, V, A):** A significant difference of MID-FiLD from earlier datasets involves the inclusion of mood annotations for fine-level dynamics. Therefore, we further examine the controllability of CC#1 values by mood. We conduct a mood recognition task using a support vector machine (SVM) to observe whether each sample can be distinguished by mood classes through a simple model (Hung et al. 2021).

Model	RMSE	Difference of min.	Difference of max.	4Q	V	A
w/o Instrument	29.4925($\pm$18.45)	20.4468($\pm$21.43)	25.5463($\pm$21.19)	0.5349	0.7698	0.6744
w/o Track role	31.9308($\pm$18.52)	25.1991($\pm$24.13)	25.7222($\pm$21.08)	0.5222	0.7307	**0.6792**
w/o Mood	28.8711($\pm$16.79)	22.1713($\pm$22.86)	27.9051($\pm$21.36)	0.4953	0.7313	0.6379
w/o Min&max	36.0331($\pm$19.37)	29.9745($\pm$25.90)	31.5162($\pm$22.64)	0.5319	0.7589	0.6714
w/o Meta	41.5318($\pm$19.34)	33.9167($\pm$27.95)	34.9074($\pm$22.49)	0.4916	0.6882	0.6595
w/ All meta	**21.6794**($\pm$17.79)	**10.9861**($\pm$15.29)	**17.2546**($\pm$19.39)	**0.5510**	**0.8058**	0.6553
Ground truth	-	-	-	0.7245	0.8495	0.8056

Table 3: Results of the ablation study excavating the impact of each element in metadata. As we reduce metadata items as inputs one by one while training, we measure each test sample's objective metric scores and calculate the mean value of all the test sample's scores in each metric. Overall, the more diverse metadata the model trains, the better fidelity and controllability the predicted result shows. Note that the difference of min/max value spiked when the model trained and predicted without minimum token and maximum token.

To this end, we extract eight features from a CC#1 sequence of each sample. These features are related to mean, standard deviation, local extremum values, and second-order gradients of the CC#1 sequence (Please refer to Appendix C.3. for more details). We use 9-fold cross-validation for training the SVM ($C = 10, \gamma = 1$) following the division ratio of the dataset. 4Q, V, and A denote the accuracy for classifying 4 quadrants, 2 valence classes, and 2 arousal classes respectively. The ground truth denotes the test set.

Results

We conducted an ablation study for objective evaluation, successively removing items in metadata as a model input. Table 3 indicates the overall experimental results regarding fidelity and controllability.

Fidelity

In Table 3, we find the best fidelity (the lowest RMSE score) in the model with all metadata and the worst in the model without any. This result is indicated by the statistical significance achieved in each score. The gradual increase of RMSE as the model excepts more items respecting metadata indicates that the five metadata have a crucial role in fidelity. When trained & predicted without min/max token, RMSE score is noticeably higher compared to other models (w/o Min&max vs. w/o Track role: $p < 0.01$), excluding the model without any metadata (w/o Meta vs. w/o Min&max: $p < 0.0001$). Although min/max metadata is mapped from extracted min/max CC#1 value, we can infer the effects on fidelity during the inference phase. Track role showed the largest influence in fidelity.

Controllability

Differences of min. & max. Comprehensively, the less metadata the model gets, the less controllability the model shows, which is also validated by statistically significant differences among the scores. The model with all metadata gets the lowest score (w/ All meta vs. w/o Instrument for both difference scores: $p < 0.0001$), proving the necessity of metadata input. On the contrary, the model without metadata gets the highest score in the same context (w/o Meta vs. w/o Min&max: $p < 0.05$). The model without min/max meta shows the second lowest (w/o Min&max vs. w/o Track role for Difference of min.: $p < 0.01$; w/o Min&max vs. w/o Mood for Difference of max.: $p < 0.05$), which demonstrates that the two tokens have an important role in controlling min/max value. The overall results say that not only min/max meta but also others have a positive effect on performance in the aspect of min/max controllability.

Mood classification. Across three classification metrics, the ground truth gets the highest accuracy scores. V scores are generally higher than the other two metrics, which indicates that determining positive or negative mood from CC#1 values can be relatively easy. In 4Q and V, the model with all metadata attains the best scores, while the model without any metadata gets the lowest scores. For classifying arousal, the model without track role shows the highest score among the models. Nonetheless, it is clear that the model without mood shows lower scores than the model with all meta regardless of the metrics. In particular, the model without mood shows the lowest A score, which is even lower than that of the model without metadata. This implies that mood annotation can be useful in estimating the right emotion with respect to valence and arousal from the fine-level dynamics in our dataset. Moreover, different tendencies in V and A scores suggest that metadata other than mood may hinder the classifier from detecting arousal.

Conclusion

In this paper, we proposed MID-FiLD, a MIDI dataset containing fine-level expressive dynamics produced by domain experts. It contains not only expressive dynamics information along with its associated notes but also metadata that can be effectively utilized as additional input for generating expressive dynamics. In addition, we demonstrated the possibility of explicit generation of expressive dynamics through token-based representation with a deep learning model. We are confident that exploration of our dataset could yield significant benefits across diverse models and further contribute to the field of music generation through appropriate and innovative representations and methodologies.

Acknowledgements

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2022 (Project Name: AI Producer: Developing technology of custom music composition, Project Number: R2022020066, Contribution Rate: 100%).

References

Berndt, A.; and Hähnel, T. 2010. Modelling musical dynamics. In *Proceedings of the 5th Audio Mostly Conference*.

Bischoff, K.; Claudiu, S.; Paiu, R.; Nejdl, W.; Laurier, C.; and Sordo, M. 2009. Music mood and theme classification - A hybrid approach. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*.

Cancino-Chacón, C. E.; Grachten, M.; Goebl, W.; and Widmer, G. 2018. Computational models of expressive music performance: A comprehensive and critical review. *Frontiers in Digital Humanities*, 5(25): 1–23.

Castellon, R.; Donahue, C.; and Liang, P. 2020. Towards realistic MIDI instrument synthesizers. In *Proceedings of the 4th Workshop on Machine Learning for Creativity and Design at NeurIPS 2020*.

Engel, J.; Hantrakul, L.; Gu, C.; and Roberts, A. 2020. DDSP: Differentiable digital signal processing. In *Proceedings of the 8th International Conference on Learning Representations*.

Gabrielsson, A.; and Juslin, P. N. 1996. Emotional expression in music performance: Between the performer's intention and the listener's experience. *Psychology of Music*, 24: 68–91.

Gomez, P.; and Danuser, B. 2007. Relationships between musical structure and psychophysiological measures of emotion. *Emotion*, 7: 377–387.

Goto, M.; Hashiguchi, H.; Nishimura, T.; and Oka, R. 2002. RWC Music Database: Popular, classical, and jazz music databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval*.

Hawthorne, C.; Stasyuk, A.; Roberts, A.; Simon, I.; Huang, C.-Z. A.; Dieleman, S.; Elsen, E.; Engel, J.; and Eck, D. 2019. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proceedings of the 7th International Conference on Learning Representations*.

Hsiao, W.-Y.; Liu, J.-Y.; Yeh, Y.-C.; and Yang, Y.-H. 2021. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Huang, C.-Z. A.; Vaswani, A.; Uszkoreit, J.; Shazeer, N.; Hawthorne, C.; Dai, A. M.; Hoffman, M. D.; and Eck, D. 2018. Music Transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*.

Huang, Y.-S.; and Yang, Y.-H. 2020. Pop Music Transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM international conference on multimedia*.

Hung, H.-T.; Ching, J.; Doh, S.; Kim, N.; Nam, J.; and Yang, Y.-H. 2021. EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*.

Jeong, D.; Kwon, T.; Kim, Y.; Lee, K.; and Nam, J. 2019a. VirtuosoNet: A hierarchical RNN-based system for modeling expressive piano performance. In *Proceedings of the 20th International Society for Music Information Retrieval*.

Jeong, D.; Kwon, T.; Kim, Y.; and Nam, J. 2019b. Graph neural network for music score data and modeling expressive piano performance. In *Proceedings of the 36th International Conference on Machine Learning*.

Jonason, N. 2020. The control-synthesis approach for making expressive and controllable neural music synthesizers. In *Proceedings of the 2020 Joint Conference on AI Music Creativity*.

Jędrzejewska, M. K.; Zjawiński, A.; and Stasiak, B. 2018. Generating musical expression of MIDI music with LSTM neural network. In *Proceedings of the 11th International Conference on Human System Interaction (HSI)*.

Kartomi, M. J. 1990. *On concepts and classifications of musical instruments*. University of Chicago Press.

Lauly, S. 2010. *Modélisation de l'interprétation des pianistes and applications d'auto-encodeurs sur des modèles temporels*. Master's thesis, University of Montréal.

Laurier, C.; Sordo, M.; Serra, J.; and Herrera, P. 2009. Music mood representations from social tags. In *Proceedings of the 10th International Society for Music Information Retrieval Conference*.

Lee, H.; Kim, T.; Kang, H.; Ki, M.; Hwang, H.; Park, K.; Han, S.; and Kim, S. J. 2022. ComMU: Dataset for combinatorial music generation. In *Proceedings of the 36th Conference on Neural Information Processing Systems*.

Levy, M.; and Sandler, M. B. 2007. A semantic space for music derived from social tags. In *Proceedings of the 8th International Society for Music Information Retrieval Conference*.

Li, B.; Liu, X.; Dinesh, K.; Duan, Z.; and Sharma, G. 2018. Creating a multi-track classical music performance dataset for multi-modal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21.

Maezawa, A. 2018. Deep piano performance rendering with conditional VAE. In *Late-Breaking Demo, the 19th International Society for Music Information Retrieval Conference*.

Maezawa, A.; Yamamoto, K.; and Fujishima, T. 2019. Rendering music performance with interpretation variations using conditional variational RNN. In *Proceedings of the 20th International Society for Music Information Retrieval Conference*.

Moog, R. A. 1986. MIDI: Musical instrument digital interface. *Journal of the Audio Engineering Society*, 34: 394–404.

Ortega, F. J. M.; Perez-Carrillo, A.; and Ramírez, R. 2019. Predicting dynamics in violin pieces with features from

melodic motifs. In *Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases*.

Phillips, S. L.; and Mace, S. 2008. Sound level measurements in music practice rooms. *Music Performance Research*, 2: 36–47.

Raffel, C. 2016. *Learning-based Methods for Comparing Sequences, with Applications to Audio-to-MIDI Alignment and Matching*. Ph.D. thesis, Columbia University.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39: 1161–1178.

Tan, H. H.; Luo, Y.-J.; and Herremans, D. 2020. Generative modeling for controllable audio synthesis of expressive piano performance. In *Proceedings of the 37th International Conference on Machine Learning*.

Thickstun, J.; Harchaoui, Z.; and Kakade, S. 2017. Learning features of music from scratch. In *Proceedings of the 5th International Conference on Learning Representations*.

Todd, N. P. M. 1992. The dynamics of dynamics: A model of musical expression. *Journal of the Acoustic Society of America*, 91(6): 3540–3550.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Proceedings of the 31st Conference on Neural Information Processing Systems*.

Weninger, F.; Eyben, F.; Schuller, B. W.; Mortillaro, M.; and Scherer, K. R. 2013. On the acoustics of emotion in audio: What speech, music, and sound have in common. *Frontiers in Psychology*, 4.

Wu, Y.; Manilow, E.; Deng, Y.; Swavely, R.; Kastner, K.; Cooijmans, T.; Courville, A.; Huang, C.-Z. A.; and Engel, J. 2022. MIDI-DDSP: Detailed control of musical performance via hierarchical modeling. In *Proceedings of The 10th International Conference on Learning Representations*.