

Masked Expressiveness: Conditioned Generation of Piano Key Striking Velocity Using Masked Language Modeling

Linzan Ye

School of Arts and Sciences, University of Rochester, New York, 14627, United States of America
lye11@u.rochester.edu

Abstract. Creating and experiencing expressive renditions of composed musical pieces are fundamental to how people engage with music. Numerous studies have focused on building computational models of musical expressiveness to gain a deeper understanding of its nature and explore potential applications. This paper examines masked language modeling (MLM) for modeling expressiveness in piano performance, specifically targeting the prediction of key striking velocity using vanilla Bidirectional Encoder Representations from Transformers (BERT). While MLM has been explored in previous studies, this work applies it in a novel direction by concentrating on the fine-grained conditioned prediction of velocity information. The results show that the model can predict masked velocity events in various contexts within an acceptable margin of error, relying solely on the pitch, timing, and velocity data encoded in Musical Instrument Digital Interface (MIDI) files. Additionally, the study employs a sequential masking and prediction approach toward rendering the velocity of unseen MIDI files and achieves more musically convincing results. This approach holds promise for developing interactive systems for expressive performance generation, such as advanced piano conducting or accompaniment systems.

Keywords: Masked Language Modeling, Music Generation, BERT.

1. Introduction

Music is a universal language that transcends linguistic and cultural barriers. The artful rendering of musical scores or ideas into audio is a crucial process that allows us to experience music in its full form. Among the various elements that performers manipulate to convey their musical intentions, dynamics and timbre are two of the most critical parameters across most musical instruments. Empirical studies have demonstrated that the perception of emotion in music is profoundly influenced by dynamics [1] and timbre [2]. Pianists, in particular, utilize varying key strike velocities to produce a wide range of timbres and dynamics [3]. Generally, faster, more forceful strikes generate a sound rich in high-frequency overtones, creating a harsher tone, while slower, more deliberate strikes yield softer, more tender tones. These dynamic and timbral variations are often combined with other piano techniques to achieve various expressive effects, such as simulating bel canto-style singing or producing shimmering, pearly running passages.

Over the past few decades, numerous scholars have explored methods for building computational models of expressiveness. One of the early examples is the KTH rule-based system [4], which models expressiveness based on musical principles used by musicians. More recently, deep learning models have increasingly been adopted to tackle this complex task. PerformanceRNN [5], for instance, is a model capable of composing and generating expressive piano music with impressive results in local phrasing and dynamics. VirtuosoNet on the other hand predicts the expressive parameters for a pre-defined score [6]. The RHEPP-Transformer represents a recent development that leverages transformer encoders for modeling expressiveness [7].

Meanwhile, the conditional controllable generation of expressive parameters has attracted growing attention from researchers. Meade et al. investigated methods to influence the output of PerformanceRNN through features such as composer name and key mode [8]. Maezawa et al. propose a model that predicts expressive parameters by conditioning on a latent space that abstractly represents a specific musical interpretation [9]. A more recent study further expands on this approach by allowing independent control over articulation, dynamics, and tempo [10].

Pianists' expressive rendering of a score requires both advanced planning and spontaneous adjustments during the performance [11]. For example, the performer must decide in advance the dynamic level at which a diminuendo will conclude. Despite extensive research in this area, relatively few studies have addressed both the planning and real-time adjustments aspects, as well as providing fine-grained control over the generated expressiveness.

PerformanceRNN, for instance, enables autoregressive real-time generation of expressive performances but does not support pre-planning of expressive parameters [5]. Previous studies address the "planning" aspect by predefining concepts or latent descriptors that guide the generation process [8-10]. However, [9] and [10] do not allow for real-time, low-level adjustments based on the notes that have already been performed. While [8] enables control changes at generation time, it relies on intermediate concepts to influence parameters, which may be less accurate and intuitive for pianists compared to direct conditioning on the notes they perform. Similar approaches to this study are seen in the work by [7] and [12], which utilizes BERT for encoding musical features but lacks explorations into achieving fine-grained conditioned generation of expressive parameters.

The author argues that a fill-in-the-blank approach to expressive parameter generation addresses both planning and adjustments according to the context, making it useful for a variety of tasks. First, by conditioning on performed note events, such a system can effectively transfer pianists' skills, potentially serving as an intuitive conducting tool that allows pianists to convey their expressive intentions more naturally. Additionally, pianists often rely on common performance strategies to express their musical ideas [13]. Building models that understand these conventions enables the system to predictably render the remaining performance details, based on a performed note sequence that carries the pianist's expressive intentions, such as phrasing, timbre, and tempo.

Therefore, this study adopts a masked language modeling approach for generating expressive parameters. Performers are expected to provide a rough draft of their intended performance, including the partial timing and velocity information. The Bidirectional Encoder Representations from Transformers (BERT) naturally became the model of choice [14]. The study focuses on a relatively simple yet effective approach: modeling only the velocity information, which plays a crucial role in shaping the dynamics and timbre of the produced piano sound.

The study makes two key contributions. First, it trains and evaluates a vanilla BERT model to capture conventional performance strategies, effectively predicting velocity events based on various surrounding contexts within a tolerable margin of error. Second, it explores a generation mode that produces more consistent and musically convincing sequences of velocity events. The results demonstrate promise for the development of advanced piano conducting and accompaniment systems, as well as for creative applications in studying the expressiveness of piano performances. The source code is available at <https://github.com/Bmoist/ExpressiveMLM>.

2. Method

2.1. The Maestro Dataset

The dataset used in this study is the Maestro dataset V3.0.0, which contains nearly 200 hours of MIDI recordings performed by virtuoso pianists on Yamaha Disklaviers during the Yamaha International Piano-e-Competition. The Disklaviers are designed to capture the nuanced details of professional piano performances in MIDI format, making this dataset ideal for modeling the intricacies of piano expressiveness. The dataset primarily captures five expressive parameters: pitch, key press time, key press velocity, release time, and pedaling. While it is true that these parameters alone cannot fully reconstruct every aspect of the actual performance, such as the pianist's physical movements, room acoustics, or the piano's exact timbre, the data is sufficiently accurate for remote professional judges to rate the performances reliably [15].

Following previous conclusion, this study hypothesizes that the choices for expressive dynamics can be predicted from factors such as register and rhythm, which are implicitly encoded in the MIDI note-on note-off events [16]. The artists' expressive timing decisions, such as tempo rubato,

accelerando, and ritardando, may also aid rather than hinder the modeling task. However, since the recordings are performed by humans, the MIDI files inevitably contain errors, such as unintended key presses. These imperfections are retained in the dataset to introduce a certain degree of variability, providing users with a more realistic context to interact with.

Lastly, due to the nature of the competition, nearly all of the performed compositions belong to the Western art music canon. This study follows the dataset's distribution and focuses on modeling expressive dynamics within this tradition.

2.2. Data Preprocessing

The study adopts the same data processing pipeline introduced in [5], as it has been proven effective in modeling expressive piano performance. The MIDI events are tokenized into four event types: note-on, note-off, time-shift, and velocity, which are then converted to one-hot encodings with a vocabulary size of 356. Pedal markings in the MIDI are converted to extended note events. Timing events are quantized to 100 steps per second, while velocity events are encoded into 32 discrete bins. Data augmentations are performed by transposing the MIDI data up and down within a major third, as well as stretching the tempo by $\pm 2.5\%$ and $\pm 5\%$. The MIDI data is then split into fragments of k tokens to fit the input length of the model.

2.3. Model Architecture

The study uses a vanilla BERT model to predict piano key strike velocity using a masked language model (MLM) training objective. The model has an input size of 128 with 256-dimensional embeddings, 4 layers, and 4 attention heads, along with learned positional encoding. Cross-entropy is used as the loss function. To predict the masked tokens, the BERT encoding output passes through a fully connected layer for classification. The study experimented with three masking mechanisms: masking velocity events only, masking all event types, and weighted masking based on event type. To differentiate velocity events from other events during training, velocity events are masked with velocity mask tokens (V), while other events (note-on, note-off, time-shift) are masked with default mask tokens (D).

There are several benefits to this approach. First, the training is self-supervised and does not require musical score or MIDI-score alignment data, allowing any valid MIDI recordings of expressive piano performances to be used with minimal preprocessing. This flexibility enables the model to be fine-tuned to a specific pianist's performance style with a few MIDI recordings from the artist, without the need for extensive additional data collection. Moreover, the fully self-supervised nature of the approach reduces the reliance on annotated data, making it more scalable and applicable to diverse datasets such as improvisations in a variety of styles.

It is important to note that although the model does not explicitly include an encoder for symbolic musical features such as staff notation and rhythm types, they can be inferred from the MIDI pitch, note-on, note-off, and velocity events. Additionally, given that professional pianists use expressive parameters to convey their interpretations of music [13], the BERT encoder is likely to learn musical encoding through these expressive parameters. Experimental results indicate that the model is responsive to the musical features and can predict conventional performance strategies, such as voicing an ascending melody line with a crescendo, balancing an Alberti bass with the main melody, and emphasizing the bass line in polyphonic passages.

2.4. Evaluation

The study evaluates the model's ability to reconstruct velocity events based on adjacent MIDI events. The model's performance is assessed using two metrics: HITS@ k (1) and accuracy within n classes (2). Formula (1), n represents the total number of classes and I is an indicator function that returns 1 if the target's rank is less than or equal to k , and 0 otherwise. In formula (2), $|C|$ stands for the total number of classes, S_c represents the set of test data with class c as their labels, and $\hat{f}(x)$ denotes the model's prediction.

$$HITS@k = \frac{1}{n} \sum_{i=1}^n I(rank_i \leq k) \quad (1)$$

$$Acc_n = \frac{1}{|C|} \sum_{c \in C} \frac{|\{y \in S_c: |\hat{f}(x) - c| \leq n\}|}{|S_c|} \quad (2)$$

Since HITS@k measures whether the top k predictions include the ground truth, it helps assess the model's ability to suggest plausible alternatives. This is particularly relevant in piano performance, where multiple valid interpretations and dynamic levels can coexist. On the other hand, due to the ordinal nature of piano dynamics, identifying adjacent velocity bins is generally more accurate than classifying distant bins. From a performance perspective, slightly off-target dynamics are preferable to sudden, unintended dynamic contrasts. Therefore, metric (2) is used to evaluate whether the top prediction falls within a specified range of the ground truth velocity bin.

3. Result

3.1. Training Details

The model was trained with a constant learning rate of 1e-4 and 10% dropout for 40 epochs. Initially, the model was trained by masking only the velocity events. However, after approximately 11 epochs, the model began to overfit, resulting in largely repetitive predictions. To address this issue, non-velocity events were also masked, forcing the model to learn both velocity and other tokens (note-on, note-off, and time-shift). These tokens implicitly carry crucial musical features that are essential for determining dynamics, ranging from low-level information such as pitch and rhythm to high-level abstractions such as expressive tempo deviations and articulations (e.g., legato, staccato). Two masking strategies were explored: masking all events with equal probability (0.15) and masking velocity events with a higher probability (0.175) than other events (0.075)

3.2. Quantitative Analysis

Including tokens other than velocity tokens aids in the training of BERT, with no signs of overfitting regardless of the masking strategy used. This suggests that the model can potentially benefit from further training. Table 1 presents the HITS@k score and accuracy within n classes (Acc±n) for the model's velocity event predictions. In the absence of relevant benchmarks, the study focuses on examining and comparing the results obtained from the two proposed masking strategies.

Table 1. HITS@k, Acc±n for predicting velocity events

	HITS@1	HITS@2	HITS@3	Acc±1	Acc±2	Acc±3
Equal Mask	0.463	0.652	0.782	0.680	0.864	0.926
Weighted Mask	0.453	0.637	0.793	0.697	0.879	0.922

Both strategies achieve a much higher HITS@1 score than random guessing (3.125%), indicating that the model is learning the relationships between velocity and surrounding musical information. Additionally, there is a significant increase in accuracy from HITS@k to Acc±n, suggesting that the model effectively learns to predict the general dynamic level based on the surrounding context, though it still struggles with achieving fine-grained precision.

When comparing the two masking strategies, the equal mask strategy achieves the highest HITS@1 and HITS@2 scores. However, the weighted masking strategy outperforms the equal mask strategy at Acc±1 and Acc±2, which are the two most critical metrics for velocity modeling. Given that there are approximately eight common dynamic levels, deviating by one bin from the ground truth equates to roughly plus or minus ¼ of a dynamic level. Although professional piano performance demands even finer dynamic control, this error range is subtle enough to serve as a promising prototype for future research. In the section below, a qualitative analysis of the generated performance is also provided, along with listening examples for readers to explore.

As a reference, the HITS@k and Acc±n results for predicting all event types are presented in Table 2. The HITS@1 and Acc±1 scores for both strategies are noticeably higher than those in Table 1,

suggesting that predicting velocity events with zero deviation is more challenging than predicting the other three events combined.

Table 2. HITS@k, Acc±n for predicting all events

	HITS@1	HITS@2	HITS@3	Acc±1	Acc±2	Acc±3
Equal Mask	0.614	0.743	0.825	0.759	0.870	0.915
Weighted Mask	0.606	0.732	0.828	0.767	0.880	0.916

Lastly, it is important to note that the metrics used do not comprehensively measure generation quality. For example, they do not evaluate the dynamic balance across voices, which is crucial in piano performance. Another important aspect is the shaping of melody lines through dynamic, which involves identifying the melody lines and comparing their phrasing, such as the location of the loudest note and the rate of crescendo/diminuendo. Moreover, since multiple valid interpretations can coexist, a model with a low accuracy score does not necessarily indicate poor performance in modeling expressive dynamics.

3.3. Case Study

The study further evaluates the model by generating a fully expressive MIDI piano performance. The initial strategy involves masking all velocity tokens and segmenting the sequence into non-overlapping chunks of 128 tokens. However, this approach presents three inherent challenges: a complete absence of dynamic level indications, missing contextual information across chunks, and an excessive number of velocity tokens to predict simultaneously, as the model is trained to predict 15% of all velocity tokens per input.

To address these challenges, the study proposes three additional generation strategies. The first strategy involves randomly leaving 10% of all velocity events unmasked to serve as a reference for the model. Additionally, overlapping chunks of tokens are created so that predictions from previous chunks can influence subsequent ones. Lastly, the model is set to predict 10% of the masked velocity events per round, allowing it to refine its predictions based on the results from previous rounds.

For clarity, the following abbreviations are used: mask all velocity tokens (MA), mask 90% of velocity tokens (M0.9), predict all tokens at once (PA), predict 10% per round (P0.1), isolated chunks (IC) and overlapped chunks (OC). Table 3 examines the effect of these strategies by comparing PA and IC with P0.1 and OC. The results are presented in terms of Acc±n.

Table 3. Acc±n for Full MIDI Performance Generation

Mask Type	Strategy	Acc±1	Acc±2	Acc±3
MA	PA, IC	0.230	0.348	0.435
	P0.1, OC	0.196	0.323	0.449
M0.9	PA, IC	0.547	0.704	0.792
	P0.1, OC	0.535	0.707	0.807

According to the results in Table 3, both strategies (PA + IC and P0.1 + OC) show a notable increase in accuracy when a random 10% of all velocity events (approximately three tokens for an input size of 128) are left unmasked, serving as clues for the model. However, the accuracy remains significantly lower compared to the results in Table 1, highlighting the complexity of this task. When comparing the two strategies, P0.1 + OC performs worse than PA + IC at Acc±1 but better at Acc±3. P0.1 + OC's performance particularly deteriorates when no contextual information is provided.

Nevertheless, as previously mentioned, the metrics alone cannot fully evaluate the model's performance. Thus, a qualitative analysis is provided. The results indicate that P0.1 + OC produces more musically convincing outputs than PA + IC. Fig. 1 displays two results generated by the PA + IC and P0.1 + OC, conditioned on a professional pianist's performance of Chopin's Op.28 Prelude No.20 in C Minor from the ATEPP dataset [17]. The screenshots are taken from Logic Pro's MIDI Piano Roll, where brighter colors represent stronger velocity.

PA + IC's outputs often include random strong velocity events that are out of context, such as the top MIDI note in the middle of the figure. Additionally, it struggles with balancing the left- and right-

hand parts, as seen in the first half of the MIDI where the right-hand part, containing the main melody, is overshadowed by random strong velocity events in the lower half notes. In contrast, P0.1 + OC's generation is smoother, contains fewer random strong velocity events, and maintains a better balance between the parts. Similar results were observed in generations conditioned on MIDI with 90% of the velocity events masked. More listening examples are available at the aforementioned GitHub link.

Interestingly, by masking all velocity events, both models consistently predicted lower dynamics in the latter half of the piece across multiple generations, corresponding to the notated *p* dynamic level in the score. Since no explicit dynamic references were provided, this may be due to how BERT encodes pitch information or perhaps the ritardando before entering the latter half. This intriguing observation warrants further investigation in future studies.

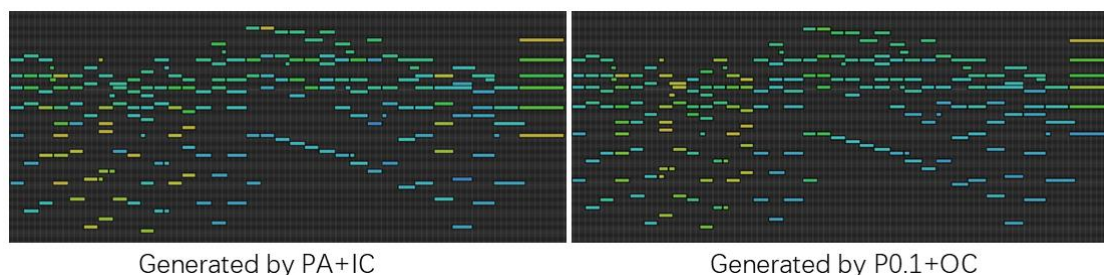


Figure 1. Representative examples generated by different methods

4. Discussion

The study investigates BERT's performance in modeling expressive piano dynamics through various generation modes and masking mechanisms. The model's efficacy is evaluated using HITS@k, Acc±n, and a case analysis. Results indicate that while BERT can effectively model the general dynamic contours of piano performance, achieving fine-grained accuracy remains challenging. However, the metrics used in the study provide only a limited analysis of generation quality. Future work should focus on developing new metrics that account for phrasing and voicing within different musical contexts. Such metrics would better capture the subjective and perceptual qualities of musical expression.

Several factors may hinder the model's ability to accurately predict the velocity of musical performances. The inherent complexities of musical expression, where subtle nuances and rich context play a critical role, pose significant challenges. The dataset may lack information on factors that velocity events depend on, such as personal style, advanced musical features, and the pianist's response to room acoustics. Additionally, the limitations could stem from the model architecture and size, as only the vanilla BERT is used in this study. Furthermore, imperfections in performances, such as mistouching unintended keys, often result in small unintended velocity events that introduce noise into the dataset. It is also possible that imperfections in the execution of a performer's expressive intentions further complicate the task of velocity prediction.

While the study concentrates on modeling key strike velocity, an expressive performance also involves intricate timing variations. Future research could extend the current model to include time-shift events, enabling it to fill in the blank with both velocity and timing deviations. Moreover, exploring the feasibility of autoregressive generation conditioned on a performance plan and user inputs in real-time applications presents an exciting challenge.

5. Conclusion

The study employs a masked language modeling (MLM) approach to predict key strike velocity events by training a vanilla BERT model. The results demonstrate that the model effectively understands the surrounding musical context and predicts velocity events within a tolerable margin of error. Additionally, the introduction of a step-by-step overlapped chunk generation strategy has

led to the generation of more coherent and musically convincing velocity markings. These findings underscore the potential of this approach to contribute to the development of new modes of interactive piano performance generation, including advanced conducting and accompaniment systems.

References

- [1] Chapin, H., Jantzen, K., Kelso, J. S., Steinberg, F., & Large, E. (2010). Dynamic emotional and neural responses to music depend on performance expression and listener experience. *PloS one*, 5 (12), e13812.
- [2] Hailstone, J. C., Omar, R., Henley, S. M., Frost, C., Kenward, M. G., & Warren, J. D. (2009). It's not what you play, it's how you play it: Timbre affects perception of emotion in music. *Quarterly journal of experimental psychology*, 62 (11), 2141-2155.
- [3] Hart, H. C., Fuller, M. W., & Lusby, W. S. (1934). A precision study of piano touch and tone. *The Journal of the Acoustical Society of America*, 6 (2), 80-94.
- [4] Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in cognitive psychology*, 2 (2), 145.
- [5] Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2020). This time with feeling: Learning expressive musical performance. *Neural Computing and Applications*, 32, 955-967.
- [6] Jeong, D., Kwon, T., Kim, Y., Lee, K., & Nam, J. (2019, November). VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance. In *ISMIR* (pp. 908-915).
- [7] Tang, J., Wiggins, G., & Fazekas, G. (2023). Reconstructing Human Expressiveness in Piano Performances with a Transformer Network. *arXiv preprint arXiv: 2306.06040*.
- [8] Meade, N., Barreyre, N., Lowe, S. C., & Oore, S. (2019). Exploring conditioning for generative music systems with human-interpretable controls. *arXiv preprint arXiv: 1907.04352*.
- [9] Fujishima, A. M. K. Y. T. (2018). Rendering music performance with interpretation variations using conditional variational rnn. *ISMIR*.
- [10] Rhyu, S., Kim, S., & Lee, K. (2022). Sketching the expression: flexible rendering of expressive piano performance with self-supervised learning. *arXiv preprint arXiv: 2208.14867*.
- [11] Woody, R. H. (1999). The relationship between explicit planning and expressive performance of dynamic variations in an aural modeling task. *Journal of research in music Education*, 47 (4), 331-342.
- [12] Chou, Y. H., Chen, I., Chang, C. J., Ching, J., & Yang, Y. H. (2021). MidiBERT-piano: large-scale pre-training for symbolic music understanding. *arXiv preprint arXiv: 2107.05223*.
- [13] Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei". *The Journal of the Acoustical Society of America*, 92 (5), 2546-2568.
- [14] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810.04805*.
- [15] Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. Z. A., Dieleman, S., ... & Eck, D. (2018). Enabling factorized piano music modeling and generation with the MAESTRO dataset. *arXiv preprint arXiv: 1810.12247*.
- [16] Rhyu, S., Kim, S., & Lee, K. (2022). Sketching the expression: flexible rendering of expressive piano performance with self-supervised learning. *arXiv preprint arXiv: 2208.14867*.
- [17] Zhang, H., Tang, J., Rafee, S., Dixon, S., Fazekas, G., & Wiggins, G. (2023). ATEPP: A dataset of automatically transcribed expressive piano performance.