Exploring the Intersection of Affect and Musical Engagement to Define Affective Computing in Musical Collaboration

Justin Adebayo Kerobo[1] and Ivica Ico Bukvic[2]

[1,2] *Institute for Creativity, Arts, and Technology, Virginia Tech, United States of America*

[1]justinkerobo@vt.edu, [2]ico@vt.edu

Abstract

This paper explores the interplay between affect and musical engagement, drawing on psychological theories and empirical studies. We propose extending this exploration into telematic music, investigating the potential for creating an emotional AI collaborator. The study aims to compile a dataset for machine learning (ML) by converting physiological responses, self-reported emotions, and musical passages. That data will allow for the generation of symbolic music from ML techniques. This exploration involves a social-psychological inquiry into human perceptions and feelings towards musical engagement and a deeper investigation into the physiological and neuroscientific responses when exposed to musical stimuli. This paper introduces Affective Computing in Musical Collaboration, exploring the intersection of physiological factors and music, particularly in telematic settings, to foster emotional expression and collaboration. It employs flow theory to understand musical engagement, examining cognitive, affective, and psychophysiological indicators to characterize flow states. Integrating AI, machine learning, and human feedback is proposed to deepen understanding and enable continuous measurement of physiological patterns, thereby advancing music information retrieval knowledge and enhancing emotional expression, collaboration, and engagement in musical performance.

Keywords: affect, musical engagement, telematic music, emotional ai collaborator, human-computer interaction

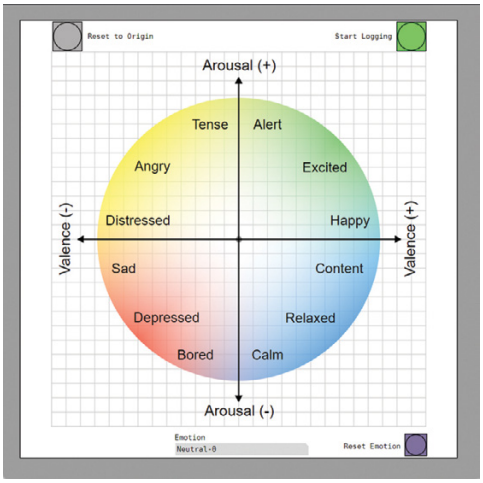
Introduction

This work defines Affective Computing in Musical Collaboration (ACMC) as a combination of physiological factors in the context of music, and extends into telematic music, resulting in an interdisciplinary field focused on developing advanced technological tools and systems to facilitate emotional expression, engagement, and collaboration in telematic music performances. Flow theory emerges as a critical framework, defining musical engagement as a state of total focus and intrinsic motivation. Various cognitive, affective, and psychophysiological indicators are examined to characterize the flow state during musical performance. Machine learning is proposed to identifty physiological flow patterns, aiming to deepen our understanding and enable continuous measurement without interrupting activities. This work advocates for interdisciplinary research integrating AI, human feedback, and machine learning to advance the knowledge of flow dynamics while defining ACMC. It highlights the importance of contextual factors in moderating flow physiology and emphasizes the need for further exploration of behavior, cognition, physiology, and flow dynamics. ACMC is essential for its ability to enhance emotional expression in music, facilitate remote collaboration, push the boundaries of human-computer interaction, advance emotional AI, as well as to enrich the user experience and engagement in musical performances. ACMC opens new possibilities for artistic expression, collaboration, and innovation in the digital age by bridging the gap between technology and musical creativity. By focusing specifically on musical collaboration, ACMC emphasizes the integration of emotion recognition technologies, music information retrieval tools, human-computer interaction principles, and AI-driven systems to enhance emotional expression, engagement, and collaboration in musical contexts.

The relationship between affect (emotional experience) and musical engagement is a complex and multifaceted topic extensively explored in psychology, neuroscience, and musicology. Various theories and studies contribute to understanding of how music evokes emotions and engages individuals both emotionally and cognitively. There are two tangents in the relationship between affect and musical engagement. A more social and psychological inference happens when querying humans about their thoughts about musical

engagement and how they feel. This is typically represented in qualitative feedback that includes more interpretations from discursive work, with deeper cultural or societal references. However, we recognize that there are strong ties to human factors and deeper connections in neuroscience. When a stimulus such as music is introduced to a human, physiological changes can be tracked along by looking at how neural coding or processing of sound frequency can be represented and measured through electroencephalography (EEG), reflecting brain stem activity that is phase-locked to sound waves and represents the sound frequency, timbre, and harmonics (Janisse, 1970; Kim et al., 2023; Wang et al., 2020; Zajonc, 1968). Physiological changes are examined using event-related potentials (ERP), which are very low voltages generated in the brain structures in response to specific events or stimuli (Blackwood & Muir, 1990; Sur & Sinha, 2009), and their amplitudes within a given time window regarding sounds or an event. For analysis, the observation is generally focused on the separability of the musical events and, in turn, the separability of ERP amplitudes. The phenomenological state posits human musical engagement—*flow*—as an active motivator and fundamental engagement metric (Csikszentmihalyi, 2009; de Manzano et al., 2010; Jackson, 2002; Wrigley & Emmerson, 2013). A range of cognitive, affective, and psychophysiological indicators during a musical performance can be used to characterize what is known as a flow state. Cognitive factors include a sense of effortlessness in action and what is sometimes referred to as total focus during those actions (Šimleša et al., 2018). Affective factors include a loss of self-consciousness and high intrinsic motivation (Landhäußer & Keller, 2012). Neural markers of flow have event-related potentials. Psychophysiological markers of flow or emotion include salivary cortisol, blood pressure, heart rate variability, and skin conductance (Nakamura & Roberts, 2016). Flow is a part of functional group psychological theory and is vital to successful group improvisations and performances. It is also an engagement metric for human-machine creative partnerships (Gifford et al., 2017; Pachet, 2006).

Therefore, while the physiological and neuropsychological methods (e.g., ECG, EEG, EMG, fMRI, eye tracking, saliva sampling, etc.) and aspects of human factors were emphasized previously in the context of flow and event-related potentials, they can broadly be understood as markers in the categories of physiology, emotion, cognition, and the contextual factor for flow. Consequently, there is a need for an overall engagement metric of flow that aligns with the current flow research results (Peifer et al., 2022). The insights gained from the recent research show a gap in clear physiological patterns of flow, indicating that this is the next major step in research. Peifer et al. (2022) argue that identifying physiological patterns of flow requires the integration of multiple indicators, suggesting that ML can help uncover these patterns in real time (Peifer et al., 2022). Moving toward that end, this argument proposes Affective Computing in Musical Collaboration (ACMC) as an interdisciplinary field that develops advanced technological tools and systems that facilitate emotional expression, engagement, and collaboration in telematic music performances. One example of such a tool is a real-



Figures 1 and 2 - L2Ork Tweeter interface (zoomed in) and Russell's circumplex of affect module in Pd-L2Ork

time, human-classified emotional MIDI dataset for symbolic music generation from the context of a computer music ensemble and through the telematic music platform, L2Ork Tweeter (Bukvic, 2020). Russell's circumplex of affect (Russell, 1980) module connects with L2Ork Tweeter to classify the emotion. It also aims to explore the integration of AI with human feedback from physiological factors and emotional context into a novice computer music ensemble (Kerobo & Bukvic, 2024). To classify the position as a further step in human-computer interaction, flow research, and machine learning with regards to human factors variables, this will first cover theoretical foundations regarding the psychology, neuroscience, and musicology perspectives from a human factors perspective and theoretical foundations regarding the machine learning, natural language processing (NLP), and symbolic music generation perspectives. Finally, the goal of investigating telematic music as a new context for emotional AI in musical collaboration has not been achieved yet. It is generally acknowledged that machine learning will advance this field of research. It is commonly recognized that reinforcement learning from human feedback (RLHF), along with various human factors, will be the way to utilize machine learning to identify and subsequently improve the flow pattern by offering suggestions or changes correlated to changes in music for emotional provocation. Furthermore, it can answer the second research question brought forth by Peifer et al., "How context conditions, such as characteristics of the task (e.g., difficulty) or conditions at the interface between context and person (e.g., task relevance), moderate the typical physiology of flow" (Peifer et al., 2022). The physiology of flow, when studied in conjunction with emotion, cognition, behavior, and contextual factors, underscores the need for further interdisciplinary research to develop AI and explore the relationships between behavior, cognition, physiology, contexts, and flow.

Extending the phenomenon

Components

The desire to contribute to music information retrieval (MIR) and flow research by studying affect (emotion) and musical engagement in the context of telematic music represents an ambitious interdisciplinary approach to understanding the complex relationship between music, emotion, and technology. This approach is represented by Kerobo and Bukvic in *Real-Time Human-Classified Emotional MIDI Dataset Integration for Symbolic Music Generation* (2024). The following subsections identify the critical components of this research.

Telematic music: It refers to music performance that involves participants in different locations connected via telecommunications networks. This allows for real-time collaboration between musicians who may be geographically dispersed, opening up new possibilities for musical interaction and exploration.

Affect and musical engagement: Affect, or emotion, plays a crucial role in how music is perceived and experienced. Understanding how unique musical elements evoke emotions and engage listeners is essential for creating impactful musical experiences. By studying affect and musical engagement in the context of telematic music, this research can gain insights into how technology-mediated interactions shape emotional responses and musical engagement.

L2Ork and L2Ork Tweeter: L2Ork (Bukvic, 2009) is a Linux-based laptop orchestra developed at Virginia Tech that enables collaborative music-making using open-source software and hardware. L2Ork Tweeter (Bukvic, 2020) serves as a tool or platform for capturing both physiological responses and self-reported emotions during musical experiences, with the Russell's circumplex of affect module and other additions. Using these tools, this research can collect data on subjective emotional experiences and physiological reactions to telematic music performances.

Encoding variables in files for datasets: MIDI files are a standard format for representing musical information in a digital format (*MIDI. Org – Expanding, Promoting, and Protecting MIDI Technology for the Benefit of Artists and Musicians around the World.*, n.d.). OpenSoundControl (OSC) is a data transport specification (an encoding) for real-time message communication between applications and hardware (Wright & Freed, 1997). OSC can be understood as a more flexible alternative to MIDI; OSC removes many of the ideological and hardware-related restrictions inherent to MIDI in favor of an open-ended user-defined address-space model that provides arbitrary parametric control through standard networking hardware (Wright & Freed, 1997). In addition, Tweeter files are defined by their own file format, which is trained on an LLM or, more recently, transformer models, capable of generating music and instruments in real-time based on parameters (Bukvic, 2020). By encoding variables such as physiological responses, self-

reported emotions, and musical passages into MIDI files, this research will create a structured dataset for analysis and training machine learning models.

Transformer for symbolic music generation: In the recent past, Recurrent Neural Networks (RNNs) were the dominant type of artificial neural network well-suited for sequential data, such as music (Eck & Schmidhuber, 2002a, 2002b; Medsker & Jain, 1999). However, the sequence-to-sequence model has issues, the main one being recurrence when the input sequence is quite long and contains a lot of information. Not every piece of the input sequence context is required at every decoding stage for all production activities (Yu et al., 2019). More recently, transformer models have revolutionized various fields, including natural language processing and image generation. Similarly, they have shown promise in symbolic music generation, offering unique advantages over traditional RNNs and other sequence modeling architectures (Sulun et al., 2022). Transformer models can be conditioned on additional information, such as genre labels, artist styles, or user preferences, to generate music that aligns with specific criteria. Transformer models can produce highly personalized and contextually relevant music compositions by incorporating conditional information into the generation process. By training a transformer on the encoded dataset, this research can develop a model capable of generating symbolic music that reflects the aesthetic preferences and emotional responses of the ensemble participating in the telematic music performance.

Personalized music generation: The ultimate goal of this research is to create a personalized music generation system that captures the unique aesthetic of the ensemble involved in the telematic music performance. The generated music can be tailored to resonate with the ensemble's preferences and emotional states by leveraging data on affect, musical engagement, and physiological responses. This interdisciplinary approach combines music, psychology, technology, and machine learning concepts to explore new frontiers in understanding and creating music. It has the potential to advance our understanding of the relationship between music and emotion and develop innovative tools for personalized music creation in collaborative settings.

Affect and musical engagement in telematic music

Understanding affect (emotion) and musical engagement in the context of telematic music is crucial for several reasons. Telematic music involves remote collaboration between musicians, often separated by geographical distances. Understanding how affect and musical engagement influence the experience of performers and audience members can help enhance the overall user experience. By catering to emotional responses and fostering engagement, telematic music performances can become more immersive and impactful. Affect plays a significant role in communication and collaboration, even in remote settings. Emotions can affect how performers interpret and respond to each other's cues, gestures, and expressions during telematic music performances. By understanding affect, musicians can better coordinate their efforts and create cohesive musical experiences across distances (Schlagowski et al., 2023). Telematic music relies on technology to facilitate remote communication and synchronization between musicians. Understanding how affect and musical engagement interact with technological interfaces can ensure the informed design of better tools and platforms for telematic music production. Technology can enhance the quality and authenticity of telematic music performances by incorporating features that support emotional expression and foster engagement. Telematic music blurs the boundaries between physical and virtual spaces, creating unique social dynamics that influence the musical interaction between participants. Studying affect and musical engagement in telematic music can illuminate these social dynamics, revealing how technology-mediated communication shapes interpersonal relationships, group dynamics, and collaborative creativity. For audiences experiencing telematic music performances remotely, understanding affect and musical engagement can facilitate a deeper connection with the music and performers. By designing experiences that evoke specific emotions and encourage active participation, telematic music can transcend geographical barriers and create meaningful connections between performers and listeners. A further extension or research proposition can utilize an audience component for emotional input and a feedback loop. Therefore, understanding affect and musical engagement in telematic music is essential for optimizing user experiences, improving communication and collaboration, informing technological design, studying social dynamics, and facilitating audience connection. This research opens new opportunities for creative expression, collaboration, and community building in the

digital age by exploring the intricate relationship between emotion, engagement, and technology in telematic music.

Affective Computing in Musical Collaboration – ACMC

Affective Computing in Musical Collaboration is an interdisciplinary field that designs technological systems that support emotional expression and engagement in telematic music-making. ACMC integrates insights from affective computing, human-computer interaction, music psychology, and machine learning to create innovative solutions that enhance the emotional and creative aspects of remote musical interactions. The components of the field are defined as follows:

Affective computing in music: ACMC leverages principles from affective computing to develop algorithms and systems capable of recognizing, interpreting, and responding to emotional cues in musical performances. This involves integrating emotion recognition technologies, such as facial expression analysis, voice analysis, and physiological sensing, into telematic music platforms, such as facial expression analysis, voice analysis, and physiological sensing.

Human-Computer Interaction (HCI) for musical collaboration: ACMC focuses on designing intuitive and responsive user interfaces that facilitate seamless collaboration between human performers and AI-driven systems. This involves developing interactive interfaces, gesture recognition systems, and adaptive feedback mechanisms that enhance the flow of communication and creativity in remote musical ensembles.

Telematic music technologies: ACMC explores novel technologies and platforms tailored explicitly for telematic music performances, including developing networked audiovisual systems, latency compensation techniques, and immersive virtual environments that enable real-time musical interaction and expression across geographical distances.

Emotional AI for music generation and interaction: ACMC advances the field of emotional AI by developing intelligent systems capable of generating expressive musical content and engaging in emotionally responsive interactions with human performers. This involves training AI models on large-scale datasets of emotional music performances and incorporating emotional intelligence algorithms into music generation and interaction frameworks.

Flow research and real-time feedback: ACMC incorporates flow research methodologies to investigate performers' psychological states during telematic music performances. By collecting qualitative and quantitative data on emotional experiences, engagement levels, and flow states, ACMC aims to provide real-time feedback to performers, enhancing their creative expression and musical communication. The research directions are as follows:

Emotional expression in telematic music: Projects investigate how emotional cues are communicated and interpreted in remote musical collaborations and how technology can facilitate expressive communication across distances.

AI-driven collaborative music making: exploring the role of AI-driven systems as collaborators and co-creators in telematic music ensembles and developing ML models that adaptively respond to human emotions and intentions.

Technological infrastructure for telematic music: designing scalable and reliable networked audiovisual systems, latency compensation techniques, and immersive environments to support telematic music performances with high emotional and creative engagement levels.

User experience design: examining user interface design principles and interaction paradigms that optimize emotional engagement, flow, and collaboration in telematic music platforms.

Ethical and sociocultural implications: This intersection examines the ethical implications of utilizing emotional AI in musical contexts and explores sociocultural factors that influence emotional expression and engagement in telematic music performances.

Pitfalls and ethical concerns in ML-mediated music: Despite its potential, AI collaboration in music poses several challenges. First, over-reliance on AI may diminish human creativity, leading to homogenized or formulaic outputs. Performers might defer expressive agency to the system, reducing improvisational spontaneity. Second, biases embedded in training datasets could reinforce narrow emotional or stylistic norms, marginalizing non-dominant musical expressions. Finally, an illusion of collaboration may arise if AI responses are misinterpreted as authentic co-creation rather than preconditioned outputs. These concerns

necessitate ethical oversight and transparent design in emotional ML systems.

Aesthetic framework for evaluation: To evaluate AI collaboration in music, this paper proposes an aesthetic framework grounded in phenomenological flow, affective depth, and perceived co-agency. Baseline comparison draws from existing telematic ensemble practices, assessing expressiveness, diversity, and coherence. This evaluation incorporates the following factors: emotional fidelity, which investigates how accurately the AI reflects the performer's intended affect; creative augmentation, which explores whether AI expands the expressive range or constrains it; flow enhancement, which examines whether immersion and engagement are sustained or disrupted by AI intervention; human-machine synergy, which investigates whether the AI is perceived as a collaborator or merely a tool. This framework promotes a reflective critique of AI's artistic role, aligning technological novelty with experiential richness. Conclusively, ACMC represents an exciting new frontier that integrates technological innovation, emotional intelligence, and artistic expression to enhance the quality and depth of telematic music collaborations. Through interdisciplinary research and collaboration, ACMC aims to advance human-computer interaction and emotional expression in musical contexts, paving the way for transformative advancements in the field.

Methodology

Tweeter dataset

The previous work starts like most music, and this continues as subjective. Pd-L2Ork and L2Ork Tweeter are used for *direct* samples to create a dataset containing pairs of MIDI files and emotional

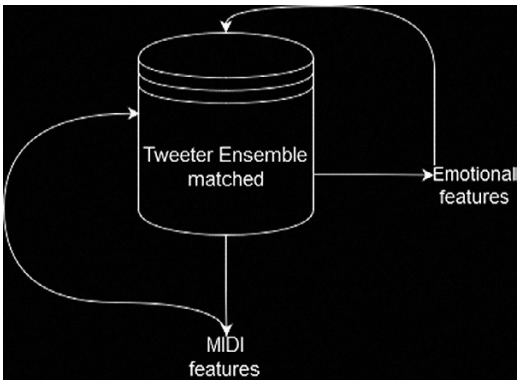


Figure 3. Dataset creation pipeline. MIDI (density, tempo, key) and emotion (valence, arousal) co-evolve in feedback loops.

labels. In particular, we match the using aesthetics and intuition as personal preferences. The reason we opted for the creation of a new dataset was to train the model on samples that are matched using a reference that is not labeled by experts, as others have done (Hung et al., 2021; Sulun et al., 2022).

Rather than correlating each MIDI sample with a song using a previous API, they are matched in real-time across the fourteen parts of L2Ork Tweeter. By doing so, the dataset creation pipeline is shortened by directly judging the MIDI information being recorded. This involves low-level MIDI features such as note density, the number of notes per second, tempo, and the correlation to the other twelve parts of the ensemble. These low-level features also model the arousal and valence dimensions of the circumplex model of affect (Russell, 1980). In totality, the existing symbolic music generation models are extended to incorporate

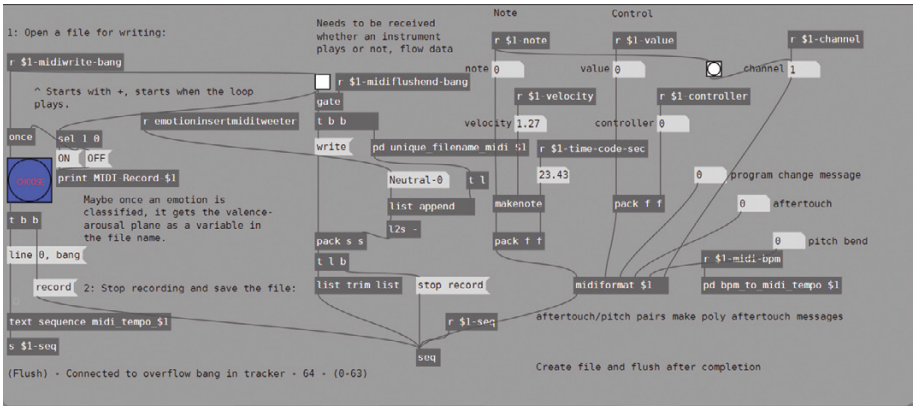


Figure 4. MIDI Subpatch. A flexible, modular system for dynamic MIDI input, enabling real-time control, visualization, storage, and complex operations like tempo mapping and file writing.

emotional conditioning and develop mechanisms to dynamically adjust model parameters based on real-time emotional input. This allows generated music to reflect the desired emotional context and implement a feedback loop where the generated music influences the performer's emotional state, creating a continuous interaction between the performer, the emotion classifier, and the music generation model.

The Pd-L2Ork MIDI subpatch in L2Ork Tweeter handles MIDI control, data display, routing, formatting, and file writing with unique filenames, using objects like receivers, float atoms, pack, and midiformat. Messages like *write* indicate commands to save MIDI data to a file using seq or sequence. For specialized MIDI handling events, the subpatch pd bpm_to_midi_tempo $1 is focused on converting beats per minute (BPM) to a MIDI tempo map message, which involves mathematical calculations to convert BPM into microseconds per beat and then further encoding it into a MIDI tempo message. For interactive elements and triggers, the bng (bang) and tgl (toggle) objects allow for user interaction, triggering various processes such as starting/stopping recordings or initiating data flow. The sel (select) and gate objects direct flow control, useful for conditional operations and data routing based on user inputs or system states. For additional elements, we have comments and instructions, like text objects. These provide instructions or annotations, guiding the user or explaining parts of the patch (e.g., "Open a file for writing," "Stop recording and save the file," etc.). Conditional logic and event handling through various trigger (t) objects are used to sequence operations, ensuring that events occur in a specific order (e.g., first clear a buffer, then start recording). Finally, there are MIDI control special cases akin to handling MIDI events like program change messages, aftertouch, and pitch bend with dedicated routes and formatting that suggest specialized use cases, possibly for performance or detailed control setups. The MIDI subpatch, combined with the Russell's circumplex of affect module, allows for synchronous writing of twelve distinctive MIDI loop-based patterns, each with its own expressive and implicit nuances. When users define the emotional state of their part or any given ensemble part, it adds that into the file name that also encodes the value of valence/arousal along with the date, time, and instrument/part it came from. By doing so, a feedback loop is created, in which a change in what users hear directly influences how they classify the emotion in correlation to the music.

Training data and pre-processing

The goal is to pre-train a non-conditional "vanilla" model for music generation on the Tweeter Dataset, which is not a subset of another database. This can be understood as any instrument, and this dataset was used to represent a new whole. It also still has separate note-on and note-off tokens for each instrument, as the polyphony is spread throughout each of the parts of the ensemble in the creation. A major difference from previous datasets (Ferreira & Whitehead, 2019; Hung et al., 2021; Panda et al., 2013; Sulun et al., 2022) and the proposed one is the fact that the material classified as a song is not a traditional song. Each of the 12 parts within L2Ork Tweeter utilizes a monophonic instrument that is coupled by a 64-note loop. The loop-based patterns are fairly small but can last up to one minute. Therefore, after pre-processing, the non-conditional training data split has 1860 loop-based patterns. To train the conditional models, the available weights from the vanilla model are first transfered and then fine-tuned on the low- and high-level for this dataset, which includes MIDI and emotional features for conditioning. After pre-processing, the conditional training data split has around 1000 loop-based patterns. Before tokenization, there is no need to convert to MIDI because the data is already in that format. We use the pretty_midi package for processing the MIDI data, as in Sulun et al. (2022). For tokenization, we use the event-based MIDI representation (Oore et al., 2020; Sulun et al., 2022). During preprocessing, we discard MIDI notes outside the piano range (21–108) and use 125 time-shift tokens (8 ms–1 s in 8 ms steps, following Oore *et al.*, 2020). Following Sulun et al. (2022), by introducing a <START> token to signify sequence commencement and a <PAD> token for sequence padding as needed, we establish the token set for our vanilla (non-conditional) model. Training input sequences are extracted as fixed-sized chunks from MIDI sequences. Each chunk's initial point is determined with a 50% probability: it either aligns with the start of a random bar, prompting insertion of the <START> token, or it is randomly selected from any point in the sequence, omitting the <START> token. This methodology proves essential for generating sequences longer than the training input length, as observed in Sulun et al.'s work (2022), during inference. A wide value range was chosen to vary loop emotions, expanding

training data. Conditioning used two values: Tweeter valence and average MIDI note density.

Models

The model's core is the music transformer, also used by Sulun et al. (Huang, Cooijmans, et al., 2019; Huang et al., 2018; Huang, Hawthorne, et al., 2019; Sulun et al., 2022). A 145M-parameter decoder-only transformer (20 layers, 768-dim, 16 heads, 3072-FFN) was used, and conditioning methods for emotion-based music generation were tested. Following Sulun et al., *discrete-token* conditioning was used by binning valence and arousal into control tokens. In thorough detail, the condition values were segmented into five equally sized bins, with the midpoint bin assigned the index 0. These bins correspond to verbal quantifiers denoting *very low*, *low*, *moderate*, *high*, and *very high* conditions, mirroring previous approaches. The control tokens associated with valence and arousal precede the music tokens, effectively concatenated along the sequence dimension, specifically when the sample originates from the commencement of a given musical bar. Subsequently, this sequence undergoes processing within the transformer architecture. An evident drawback of this model surfaces during inference. Once the generated sequence reaches the same length as the input, the input is truncated from the start, resulting in the omission of control tokens. Additionally, information loss is incurred due to the discretization of continuous values. Adhering to precedent, normalized continuous condition values for the *continuous-token* input are utilized. Each value undergoes processing through an individual linear layer, yielding condition vectors of equivalent length to the music token embeddings. These vectors and embeddings are then concatenated along the sequence dimension and passed into the transformer. Even throughout inference, including after the generated length aligns with the input length, we persist in inserting the condition vectors at the sequence's outset. Building upon this foundation, the final approach, termed *continuous-concatenated*, entails consolidating the two normalized continuous condition values into a single vector. This vector is replicated along the sequence dimension and concatenated with each music token embedding. The conditioning vectors and token embeddings maintain lengths of 192 and 576, respectively, ensuring consistency in the total feature length of the transformer input across all models. All conditional models undergo training via fine-tuning of the pre-trained vanilla (non-conditional) model. A representation of the models is not in the focus of this project but it can be found in Sulun et al. (2022).

Evaluation

Assessing music generation models remains a dynamic field, with no prevailing consensus currently established. Rather than embarking on intricate endeavors to replicate subjective listening experiments, this study employs objective, quantitative evaluation methods. The present model evaluation employs negative log-likelihood (NLL), top-1, and top-5 accuracies as metrics, following the methodology outlined by Hung et al. (2021) and Sulun et al. (2022). In assessing top-*n* accuracy, a model's prediction is deemed accurate if the ground-truth class ranks within the top n probabilities of the model's output for each token. Consistent with this training setup, evaluation is conducted on chunks of length 1216, with loss calculated for every token in the target sequence. This approach presents a heightened challenge compared to predicting the subsequent token in a sequence, particularly evident when the model attempts to predict the initial note of a song solely from the <START> token. To ensure comprehensive coverage of the test split, non-overlapping chunks are processed sequentially, yielding 1836 for evaluation. Additionally, a quantitative assessment of emotional content in samples generated by these conditional models is undertaken, as previously explored by Hung et al. (2021) and Sulun et al. (2022). To achieve this, a regression model is trained on the training data split to predict emotion values. This model, structured as a music transformer with eight layers, produces two continuous values: valence and arousal. Subsequently, employing the trained conditional generation models, inference is conducted using various conditions and emotional content is predicted using the regression model. The normalized L_1-distance is utilized between the regression model predictions and the conditions provided during inference for error quantification. To ensure equitable comparison with the discrete-token model, condition values are selected as the midpoints of the bins corresponding to the discrete condition tokens (-0.8, -0.4, 0, 0.4, and 0.8). The present paper mirrors Sulun et al. (2022) by generating a set of 25 condition value pairs for evaluation by employing five values for valence and arousal each. Eight samples are generated without selective bias for each model-condition combination and the average error is reported.

Each sample consists of 4096 tokens, and the regression model processes inputs of length 1216, mirroring the generator setup. Sample inputs for the regression model utilize a sliding window with 50% overlap, and the outputs are subsequently averaged.

Results

Table 1 presents the performance of various models based on prediction accuracy. The continuous-concatenated model demonstrates superior performance across all metrics compared to other models, notably surpassing the state-of-the-art discrete-token model by a significant margin, particularly in negative log-likelihood and top-1 accuracy. In Table 2, the regression-based evaluation further confirms the continuous-concatenated model's superiority in conveying emotion compared to alternative models. Notably, the vanilla approach is omitted from this analysis as it lacks emotional conditioning.

Tables 1 and 2. Table 1 (T1): Model performance was evaluated using NLL (lower is better), Top-1/ Top-5 accuracy (higher is better). Table 2 (T2): 'Error' (normalized L1 distance between inference conditions and regression outputs).

(T1) Model	NLL	(T2) Model	Error
vanilla	0.7856	discrete-token	0.2305
discrete-token	0.7546	continuous-token	0.2166
continuous-token	0.7387	continuous-concatenated	**0.2162**
continuous-concatenated	**0.7101**		
Top-1	Top-5		
0.7992	0.9364		
0.8061	0.9379		
0.8152	0.9385		
0.8208	0.9388		

Analyzing the performance gap among the models presented, the primary limitation of discrete-token and continuous-token models is attributed to their treatment of condition values. Unlike the continuous-concatenated model, these models assign equal importance to condition values and tokens in the sequence. Sulun et al. (2022) argue that while tokens aid local predictions, condition values have a global impact, directly influencing the entire generated sample. Their proposed continuous-concatenated model effectively leverages condition information by integrating it into each embedding of the input sequence for the transformer. The present models are slightly inferior to the former, with the unique dataset deriving minimal loss compared to larger datasets.

Subsequent exploration of these methods can incorporate continuous-valued conditions, offering finer control over the generation process. This approach also enables dynamic adjustments to conditions throughout the generation, potentially yielding more intricate and progressive compositions. In summary, this work represents a significant step towards establishing a clearer connection between emotion, symbolic music, and collaboration over distance (telematics). The Tweeter dataset currently exhibits minimal loss in comparison to the Lakh-Spotify dataset, and the individual personalizes the representations. In this case, it does this with a considerably smaller dataset, but it will be improved soon by more human-generated loop-based patterns.

Conclusion

Understanding affect and musical engagement in telematic music is crucial for several reasons. It allows for a deeper comprehension of how emotions shape the experience of both performers and audience members in remote musical collaborations, enriching user experiences and fostering meaningful connections despite geographical distances. Additionally, studying affect and musical engagement can enhance human-computer interaction in music-making processes. Advancements in emotional ML influenced by human factors in music can lead to more intuitive and responsive technology that complements human creativity and expression. Integrating emotional AI into telematic music via L2Ork Tweeter facilitates seamless collaboration and opens new avenues for exploring the emotional dimensions of music through technology. Furthermore, tracking affect and musical engagement through qualitative and quantitative data enables real-time feedback mechanisms, offering insights into the dynamic interplay between performers, audience, and technology. This real-time feedback loop contributes to the next dimension of flow research, where performers can achieve heightened focus, immersion, and creativity. By encoding variables such as physiological responses, self-reported emotions, and musical passages into

MIDI files, this approach bridges the gap between technology and musical creativity, paving the way for innovative approaches to music information retrieval. Affective Computing in Musical Collaboration (ACMC) represents an interdisciplinary approach to leveraging technology to enhance emotional expression, engagement, and collaboration in musical contexts, particularly in telematic music performances. By integrating insights from emotion recognition, human-computer interaction, ML, telematics, user experience design, and ethics, ACMC aims to push the boundaries of what is possible in the intersection of emotion, technology, and music, enriching the experiences of performers and audiences alike. In conclusion, understanding affect and musical engagement in telematic music enhances user experiences and human-computer interaction and fuels advancements in emotional AI, music information retrieval, and flow research. By integrating qualitative and quantitative data tracking and encoding variables into MIDI files through L2Ork Tweeter (Kerobo & Bukvic, 2024), a telematic music environment, we embark on a journey to deepen our understanding of music's emotional and creative dimensions, ultimately enriching the intersection of technology and musical expression through human factors, HCI through telematic music, and MIR for machine learning.

References

Blackwood, D. H. R., & Muir, W. J. (1990). Cognitive Brain Potentials and their Application. *The British Journal of Psychiatry*, 157(S9), 96–101. https://doi.org/10.1192/S0007125000291897

Bukvic, I. (2009, January 1). *L2Ork » Linux Laptop Orchestra.* https://l2ork.bukvic.net/main/

Bukvic, I. (2020, January 1). *L2Ork » L2Ork Tweeter.* https://l2ork.bukvic.net/main/make-your-own-l2ork/tweeter/

Csikszentmihalyi, M. (2009). *Flow: The Psychology of Optimal Experience*. Harper Collins.

de Manzano, Ö., Theorell, T., Harmat, L., & Ullén, F. (2010). The psychophysiology of flow during piano playing. *Emotion, 10*(3), 301–311. https://doi.org/10.1037/a0018432

Eck, D., & Schmidhuber, J. (2002a). *A First Look at Music Composition using LSTM Recurrent Neural Networks* [Technical Report]. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale.

Eck, D., & Schmidhuber, J. (2002b). Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. In *Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing* (pp. 747–756). https://doi.org/10.1109/NNSP.2002.1030094

Ferreira, L., & Whitehead, J. (2019). Learning to Generate Music with Sentiment. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*. http://archives.ismir.net/ismir2019/paper/000045.pdf

Gifford, T., Knotts, S., Kalonaris, S., & McCormack, J. (2017). Evaluating Improvisational Interfaces. In *Proceedings of the Improvisational Creativity Workshop 2017*. https://www.semanticscholar.org/paper/Evaluating-Improvisational-Interfaces-Gifford-Knotts/cc4f195806907b7be79cff4d3e4708e2b9fa8c2e

Huang, C.-Z. A., Cooijmans, T., Roberts, A., Courville, A., & Eck, D. (2019). *Counterpoint by Convolution* (No. arXiv:1903.07227). arXiv. https://doi.org/10.48550/arXiv.1903.07227

Huang, C.-Z. A., Hawthorne, C., Roberts, A., Dinculescu, M., Wexler, J., Hong, L., & Howcroft, J. (2019). *The Bach Doodle: Approachable music composition with machine learning at scale* (No. arXiv:1907.06637). arXiv. https://doi.org/10.48550/arXiv.1907.06637

Huang, C.-Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018). *Music Transformer*. arXiv:1809.04281 [Cs, Eess, Stat]. http://arxiv.org/abs/1809.04281

Hung, H.-T., Ching, J., Doh, S., Kim, N., Nam, J., & Yang, Y.-H. (2021). EMOPIA: *A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation* (No. arXiv:2108.01374). arXiv. https://doi.org/10.48550/arXiv.2108.01374

Jackson, H. (2002). Chapter 11—Toward a Symbiotic Coevolutionary Approach to Architecture. In P. J. Bentley & D. W. Corne (Eds.), *Creative Evolutionary Systems* (pp. 299–313). Morgan Kaufmann. https://doi.org/10.1016/B978-155860673-9/50049-5

Janisse, M. P. (1970). Attitudinal effects of mere exposure: A replication and extension. *Psychonomic Science, 19*(2), 77–78. https://doi.org/10.3758/BF03337428

Kerobo, J. A., & Bukvic, I. I. (2024). Real-Time Human-Classified Emotional MIDI Dataset Integration for Symbolic Music Generation. *Proceedings of the 2024 International Conference on Machine Learning and Applications (ICMLA)* (pp. 520–527). https://doi.org/10.1109/ICMLA61862.2024.00076

Kim, T., Chung, M., Jeong, E., Cho, Y. S., Kwon, O.-S., & Kim, S.-P. (2023). Cortical representation of musical pitch in event-related potentials. *Biomedical*

Engineering Letters, 13(3), 441–454. https://doi.org/10.1007/s13534-023-00274-y

Landhäußer, A., & Keller, J. (2012). Flow and Its Affective, Cognitive, and Performance-Related Consequences. In S. Engeser (Ed.), *Advances in Flow Research* (pp. 65–85). Springer. https://doi.org/10.1007/978-1-4614-2359-1_4

Medsker, L., & Jain, L. C. (1999). *Recurrent neural networks: Design and applications*. CRC Press.

MIDI.org – Expanding, promoting, and protecting MIDI technology for the benefit of artists and musicians around the world. (n.d.). Retrieved October 5, 2024, from https://midi.org/

Nakamura, J., & Roberts, S. (2016). The Hypo-egoic Component of Flow. In K. W. Brown & M. R. Leary (Eds.), *The Oxford Handbook of Hypo-egoic Phenomena*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199328079.013.9

Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2020). This time with feeling: Learning expressive musical performance. *Neural Computing and Applications, 32*(4), 955–967. https://doi.org/10.1007/s00521-018-3758-9

Pachet, F. (2006). Enhancing individual creativity with interactive musical reflexive systems. In I. Deliège & G. A. Wiggins (Eds.), *Musical Creativity* (pp. 375–391). Psychology Press. https://doi.org/10.4324/9780203088111-35

Panda, R. E. S., Malheiro, R., Rocha, B., Oliveira, A. P., & Paiva, R. P. (2013). Multi-Modal Music Emotion Recognition: A New Dataset, Methodology and Comparative Analysis. In *Proceedings of the 10th International Symposium on Computer Music Multidisciplinary Research (CMMR 2013)*, (570–582). https://mir.dei.uc.pt/pdf/Conferences/MOODetector/CMMR2013_MultiModal.pdf

Peifer, C., Wolters, G., Harmat, L., Heutte, J., Tan, J., Freire, T., Tavares, D., Fonte, C., Andersen, F. O., van den Hout, J., Šimleša, M., Pola, L., Ceja, L., & Triberti, S. (2022). A Scoping Review of Flow Research. *Frontiers in Psychology, 13*, 815665. https://doi.org/10.3389/fpsyg.2022.815665

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Schlagowski, R., Nazarenko, D., Can, Y., Gupta, K., Mertes, S., Billinghurst, M., & André, E. (2023). Wish You Were Here: Mental and Physiological Effects of Remote Music Collaboration in Mixed Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). https://doi.org/10.1145/3544548.3581162

Šimleša, M., Guegan, J., Blanchard, E., Tarpin-Bernard, F., & Buisine, S. (2018). The Flow Engine Framework: A Cognitive Model of Optimal Human Experience. *Europe's Journal of Psychology, 14*(1), 232–253. https://doi.org/10.5964/ejop.v14i1.1370

Sulun, S., Davies, M. E. P., & Viana, P. (2022). Symbolic Music Generation Conditioned on Continuous-Valued Emotions. *IEEE Access, 10*, 44617–44626. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3169744

Sur, S., & Sinha, V. K. (2009). Event-related potential: An overview. *Industrial Psychiatry Journal, 18*(1), 70–73. https://doi.org/10.4103/0972-6748.57865

Wang, S., Wang, T., Chen, N., & Luo, J. (2020). The preconditions and event-related potentials correlates of flow experience in an educational context. *Learning and Motivation, 72*, 101678. https://doi.org/10.1016/j.lmot.2020.101678

Wright, M. J., & Freed, A. (1997). Open SoundControl: A New Protocol for Communicating with Sound Synthesizers. In *Proceedings of the International Conference on Mathematics and Computing*. http://hdl.handle.net/2027/spo.bbp2372.1997.033

Wrigley, W. J., & Emmerson, S. B. (2013). The experience of the flow state in live music performance. *Psychology of Music, 41*(3), 292–305. https://doi.org/10.1177/0305735611425903

Yu, Y., Si, X., Hu, C., & Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. *Neural Computation, 31*(7), 1235–1270. https://doi.org/10.1162/neco_a_01199

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*(2, Pt.2), 1–27. https://doi.org/10.1037/h0025848

https://doi.org/10.17234/9789533793085.26