POLITECNICO DI TORINO
Repository ISTITUZIONALE

SCHuBERT: A Real-Time End-to-End Model for Piano Music Emotion Recognition.

Original
SCHuBERT: A Real-Time End-to-End Model for Piano Music Emotion Recognition / Zanoni, M., Rossi, R., Sironi, A., Belluco, P., Rottondi, C.. - ELETTRONICO. - (2026). (6th International Conference on Audio for Virtual and Augmented Reality and Immersive Games Parigi (Fra) 30 giugno-3 luglio 2026).

Availability:
This version is available at: 11583/3017042 since: 2026-10-26T19:25:40Z

Publisher:
Audio Engineering Society

Published
DOI:

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

19 November 2026

Audio Engineering Society

Conference Paper

Presented at the AES 2026 International Conference on
Audio for Virtual and Augmented Reality and Immersive Games
2026 June 30 – July 3, Sorbonne University/d'Alembert & IRCAM/STMS, Paris, France

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

SCHuBERT: a real-time end-to-end model for piano music emotion recognition

Massimiliano Zanoni[1], Riccardo Rossi[2], Alice Sironi[1], Paolo Belluco[3], Cristina Rottondi[1]

[1] *Politecnico di Torino, Italy* [2] *Politecnico di Milano, Italy* [3] *LWT3 srl, Italy*

Correspondence should be addressed to Massimiliano Zanoni (massimiliano.zanoni@polito.it)

ABSTRACT

In this study, we present SCHuBERT, a real-time end-to-end Piano Music Emotion Recognition (PMER) system that operates directly on raw audio and fine-tunes DistilHuBERT for short-window classification on the Valence–Arousal plane. Designed for low latency and high responsiveness, the system is particularly well suited for immersive applications such as virtual and augmented reality. Compared with both audio- and symbolic-domain baselines, SCHuBERT achieves strong accuracy in four-quadrant classification as well as in binary arousal and valence tasks, while maintaining low computational overhead for real-time operation.

1 Introduction

Immersive audio technologies are rapidly transforming the way music is created, experienced, and shared. Advances in Virtual Reality (VR), Augmented Reality (AR), and interactive game audio have enabled the emergence of virtual environments in which sound is no longer a passive element but becomes an active component of user interaction and perception. In these contexts, audio systems are expected to be adaptive, responsive, and tightly integrated with user behavior, spatial rendering, and multimodal feedback. This shift toward interactive and immersive sound design introduces new challenges for audio analysis and synthesis, particularly in enabling systems that can interpret expressive content and react to it in real time.

The analysis of expressive elements of performance is particularly relevant in the context of live music, across a wide range of applications [1] [2]. Such importance becomes even more significant in VR systems, where expressivity plays a key role both in enhancing interactive experiences with audiences and/or other musicians—often limited by the lack of physical presence and embodiment [3] [4]—and in extending the expressive capabilities of the experience itself, for instance through the automatic and interactive generation of virtual environments [5][6]. Among the various expressive dimensions investigated in the literature, perceived emotion stands out as one of the most powerful and extensively studied, due to its central role in shaping user experience and engagement.

Music Emotion Recognition (MER) is a branch of Music Information Retrieval (MIR) aimed at enabling machines to interpret and respond to human affective states. In recent years, MER has attracted increasing attention in interactive and adaptive musical systems, where the ability to estimate emotional content in real time can support responsive and affect-aware human–computer interaction. Emerging scenarios such

as Networked Music Performance [7], Virtual and Augmented Reality (VR/AR), and the Musical Metaverse [8], which envision virtual spaces where users can participate in live musical performances as both performers and audiences, further highlight the need for low-latency emotion recognition techniques capable of operating continuously on live musical signals, and open new possibilities for affect-aware interaction and co-creation.

Despite significant advances in deep learning for audio analysis, most MER approaches are designed for offline processing and do not explicitly address the constraints of real-time operation. However, interactive musical scenarios require systems capable of continuous low-latency inference while maintaining responsiveness and temporal consistency across streaming audio inputs [9]. Real-time MER therefore introduces several challenges beyond accurate emotion classification, including computational efficiency, streaming compatibility, and robustness under short analysis windows.

Furthermore, the inherently subjective and context-dependent nature of emotional perception complicates the design of robust and generalizable models, particularly in multimodal and interactive environments where audio may interact with visual, spatial, and social cues. These considerations motivate the development of MER systems specifically designed for real-time operation, capable of combining accurate affective analysis with low-latency processing suitable for interactive applications.

Within this framework, this paper focuses on Piano Music Emotion Recognition (PMER) as a case study for affect-aware musical interaction in immersive environments. We propose *SCHuBERT* (Sentiment Classifier Hidden-unit BERT), an end-to-end, raw-audio model designed for real-time PMER in interactive and metaverse-oriented contexts. Unlike traditional approaches relying on handcrafted features or symbolic (MIDI) representations, SCHuBERT processes raw audio waveforms directly, enabling the model to learn data-driven representations aligned with perceptually relevant emotional cues. The architecture builds on DistilHuBERT [10], a distilled self-supervised model derived from HuBERT [11], providing efficient and high-quality latent audio embeddings suitable for deployment in latency-sensitive applications. SCHuBERT demonstrates particular suitability for feature extraction due to its capacity to capture complex acoustic representations through holistic analysis of entire audio sequences. To meet the requirements of immersive performance systems, SCHuBERT incorporates a lightweight, concurrent streaming pipeline that supports real-time inference and integration within game engines and XR audio frameworks.

The main contributions of this work are: (*i*) an end-to-end PMER model that fine-tunes DistilHuBERT on short audio windows for four-quadrant emotion classification based on the valence-arousal (V-A) representation of emotion in Russell's circumplex model [12], enabling continuous affect tracking; (*ii*) a low-latency, concurrent streaming architecture designed for real-time operation in interactive and immersive audio environments; and (*iii*) an evaluation of SCHuBERT on the EMOPIA [13] dataset, with a particular focus on performance under real-time constraints. The proposed real-time variant outperforms the offline baseline while ensuring low-latency inference, making it suitable for interactive and immersive applications.

2 Related Work

The analysis of musical expressivity and perceived emotions plays a key role in applications such as NMP and interactive or generative VR systems, where it supports expressive communication, adaptive content generation, and immersive user experiences despite spatial or temporal constraints [8][9]. It is worth mentioning the work by Forero *et al.* [6] that explores affect-aware immersive environments, such as Emotional Machines, which leverage speech emotion recognition to drive the real-time generation of audiovisual virtual spaces, demonstrating how emotional states can be encoded and experienced within interactive 3D environments.

A common foundation for MER is the V–A representation [12], which captures the continuous nature of emotional perception in music. Different models combining V-A plane and English affective terms have been proposed [14]. Early approaches relied on handcrafted acoustic descriptors and classical machine-learning techniques: for instance, Yang *et al.* [15] proposed a regression-based framework mapping audio features to affective dimensions, while Lu *et al.* [16] introduced methods for automatic mood detection and tracking in audio signals. These approaches established the basis for MER but depend heavily on feature engineering and are typically designed for offline analysis.

Subsequent work explored temporal modeling of emotion, moving toward continuous prediction. Weninger *et al.* [17] proposed an online framework using recurrent neural networks for continuous-time mood regression, while Li *et al.* [18] introduced multi-scale fusion strategies for dynamic emotion prediction. These methods are conceptually related to our work in modeling temporal affect evolution; however, they rely on engineered features and are not optimized for low-latency streaming from raw audio.

More recent approaches leverage deep neural networks to improve MER performance. Architectures based on convolutional and recurrent networks [19, 20, 21] and attention mechanisms [22] have demonstrated improved ability to capture complex musical patterns. While these models share with SCHuBERT the use of deep learning for representation learning, they are generally designed for offline processing and often rely on time-frequency representations rather than raw waveform input.

In the context of PMER, the EMOPIA dataset [13] provides a key benchmark with aligned piano performances and emotion annotations. On the symbolic side, MidiBERT-Piano [23] showed that large-scale transformer pre-training on MIDI data can significantly improve emotion-related tasks. However, symbolic approaches inherently lack performance-level acoustic cues such as timbre, articulation, and expressive nuances. On the audio side, Wang *et al.* [24] demonstrated that Mel-based convolutional networks can effectively model emotional content in piano music. Recent work by Zheng and Li [25] further explores real-time emotion-aware piano music generation using GAN-based architectures, highlighting the growing interest in integrating affect modeling within interactive and generative musical systems. SCHuBERT is analogous in its use of acoustic input, but differs by adopting an end-to-end waveform-based approach that eliminates manual feature design.

Multimodal approaches have also been explored to combine complementary information sources. Zhao and Yoshii [26] proposed a self-attentive fusion of symbolic and acoustic features for MER. While such approaches improve performance, they assume the availability of multiple modalities and increase system complexity. In contrast, SCHuBERT focuses on a raw-audio-only formulation, which is more suitable for real-time deployment in performance scenarios where symbolic data may not be available.

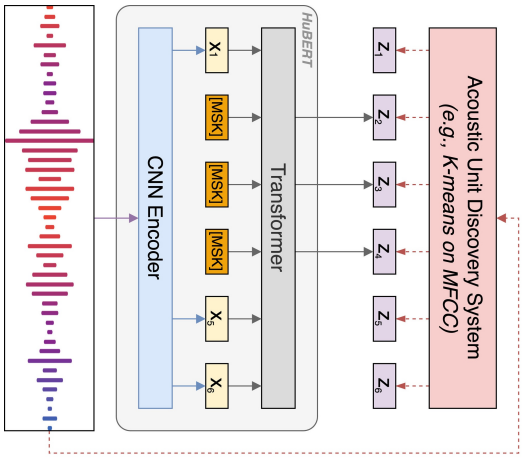


Fig. 1: HuBERT architecture: a CNN encoder and Transformer predict masked frames using discrete targets from an acoustic unit discovery system.

The representation-learning backbone of our method builds on recent advances in self-supervised audio modeling. HuBERT [11] demonstrated that masked prediction of latent units enables the learning of transferable audio representations, while DistilHuBERT [10] introduced a more compact model through knowledge distillation, making it suitable for latency-sensitive applications. SCHuBERT extends these approaches by adapting a distilled self-supervised encoder to PMER and embedding it within a concurrent streaming pipeline for real-time inference.

3 The HuBERT backbone

Hidden-Unit BERT (HuBERT) [11] is a self-supervised, BERT-like model designed for learning meaningful representations from raw audio. It leverages masked prediction to jointly model acoustic features and temporal dependencies (Fig. 1).

Given a sequence of audio frames $X = [x_1, \ldots, x_T]$, HuBERT associates each frame with a discrete target label $Z = [z_1, \ldots, z_T]$ obtained via offline clustering. Specifically, acoustic features extracted from the audio are clustered using unsupervised algorithms. Clustering algorithms group data samples according to their similarity in the feature space, without relying on manually annotated labels. In this context, they allow acoustically similar frames to be associated with the same discrete category, capturing recurring patterns in the audio

representation. The resulting cluster assignments are then used as pseudo labels during pre-training. These discrete targets provide a form of self-supervision that allows the model to learn meaningful latent acoustic representations. Training is performed by masking a subset of frames M and predicting their corresponding targets. The objective combines masked and unmasked prediction losses:

$$L = \alpha L_m + (1-\alpha) L_u, \tag{1}$$

where L_m is computed over masked timesteps and L_u over unmasked ones.

To improve target quality, HuBERT employs *cluster ensembles*, combining multiple clustering models to provide complementary supervision at different granularities.

DistilHuBERT [10] extends HuBERT through knowledge distillation, transferring representations from a large teacher model to a smaller student. The student is trained to match intermediate hidden representations of selected teacher layers using a combination of L1 and cosine similarity losses. Lets consider the D dimensional feature vectors at time t^{th} generated by the teacher's l^{th} layer and the student's corresponding prediction head respectively as $h_t^{(l)}$ and $\hat{h}_t^{(l)}$:

$$L^{(l)} = \sum_{t=1}^{T} \left(\frac{1}{D} \left\| h_t^{(l)} - \hat{h}_t^{(l)} \right\|_1 - \lambda \log \sigma(\cos(h_t^{(l)}, \hat{h}_t^{(l)})) \right), \tag{2}$$

where σ is the sigmoid activation function and $\lambda > 0$ controls the contribution of the cosine similarity loss.

The resulting model retains most of HuBERT's representational power while significantly reducing computational cost (up to $\sim$73% speedup with respect to HuBERT), making it well suited for real-time applications.

4 SCHuBERT

The methodology proposed in this study, SCHuBERT, leverages HuBERT [11] to extract meaningful latent representations from music audio and adapts them to PMER on short raw audio excerpts. Specifically, the model fine-tunes DistilHuBERT [10], a distilled and computationally efficient version of HuBERT that retains its representational capabilities while reducing inference cost. Pre-trained via knowledge distillation on hidden representations, DistilHuBERT enables the extraction of compact and expressive features, making it suitable for real-time applications.

SCHuBERT adopts an end-to-end architecture that directly maps raw audio waveforms to the target output, eliminating the need for handcrafted feature extraction. During training, the model automatically learns task-relevant representations from the input signal. This approach is particularly advantageous for real-time applications, where low-latency processing is essential. Furthermore, by operating on raw audio instead of symbolic representations such as MIDI, SCHuBERT preserves the full acoustic complexity of live piano performances, including timbral nuances, dynamics, articulation, and expressive characteristics.

Fig. 2 illustrates the overall pipeline of the proposed method. The initial MIDI-to-audio conversion stage is optional and is applied only when audio recordings are not directly available. The input audio is then segmented into short excerpts using a windowing procedure parameterized by the *window size* (w) and *hop length* (h). Each segment is subsequently processed through a *peak normalization* step (target peak equal to 1.0), which scales the signal amplitude so that its maximum reaches a fixed reference level, ensuring consistent loudness across different recordings.

DistilHuBERT is employed to extract high-level representations from each short audio segment, yielding compact latent embeddings. These representations are then used for real-time emotion classification through a fine-tuning stage built on top of the DistilHuBERT backbone. Specifically, a lightweight classifier is appended, consisting of a linear layer, a dropout layer for regularization, and a final softmax layer. The last hidden states of DistilHuBERT are pooled into a fixed-length one-dimensional feature vector (size 768), which is passed through the classifier. The softmax layer maps the output to four neurons, corresponding to the target emotion classes

It is worth noting that, in live piano performances, the system must continuously acquire and process audio in real time, segmenting the incoming stream into fixed-length excerpts for emotion recognition. The fundamental challenge lies in ensuring uninterrupted audio capture while simultaneously performing computationally intensive inference. This requirement introduces a critical trade-off between processing time and data

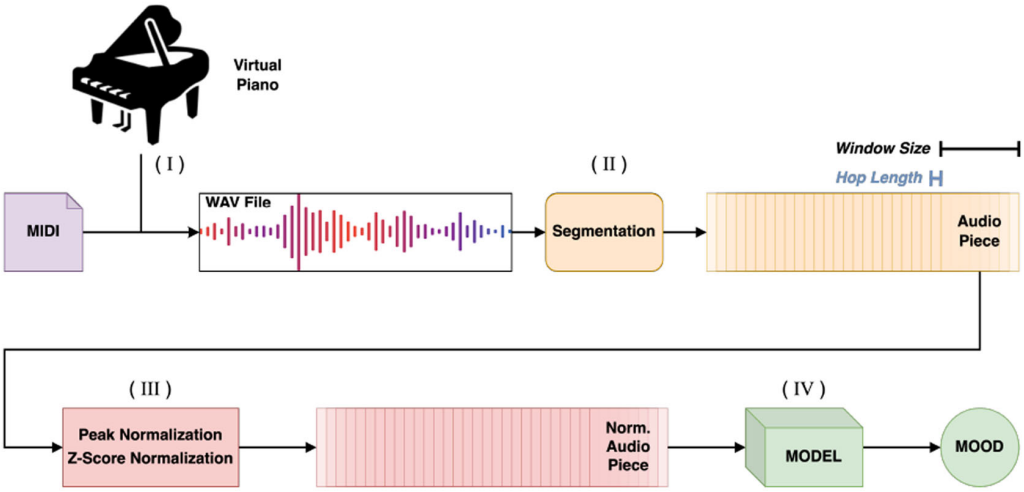


Fig. 2: SCHuBERT pipeline: MIDI is converted to audio, segmented into overlapping windows, normalized, and processed by the model to predict the emotional state.

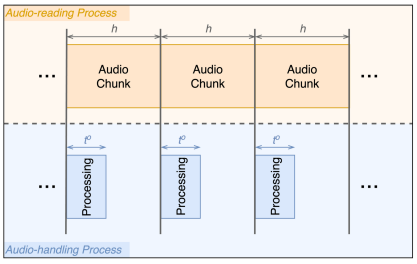


Fig. 3: Parallel audio processing: incoming audio is acquired in chunks of size h while processing is performed concurrently on sliding windows w, ensuring continuous real-time operation without interruptions.

acquisition, as any delay in computation may interfere with the continuity of the input stream, potentially leading to data loss or perceptible latency.

A key challenge in real-time audio processing arises from the inherently sequential nature of standard computing pipelines, which require that the system must first acquire an entire audio segment of size w, then process it, before proceeding to the next acquisition. In this context, the adopted streaming approach leverages the hop length h to incrementally acquire the audio signal. At each step, a new chunk of duration h is captured and appended to a buffer containing the most recent past samples, so as to construct a complete analysis window of size w. Emotion recognition is then performed on this aggregated segment, which combines the newly acquired audio with the historical context. This mechanism enables continuous updates of the emotional state while preserving temporal coherence, without requiring the system to wait for entirely new, non-overlapping windows. Under typical operating conditions, using a single process (or thread) for both audio capture and processing would disrupt the continuity of the audio stream. Specifically, after each chunk h is acquired, the process would need to suspend input acquisition to process the accumulated window w, leading to temporal discontinuities.To overcome this limitation, we design a parallel architecture based on two concurrent processes. The first process is dedicated to continuous audio acquisition, while the second process receives the streamed chunks and performs emotion recognition once the required window size is reached (see Fig. 3). This design guarantees uninterrupted audio capture during live performances while preserving real-time responsiveness.

5 Experimental setup and evaluation

5.1 Affective Model

Emotions in music can be modeled through either discrete emotional categories or continuous dimensional representations. Among the dimensional approaches,

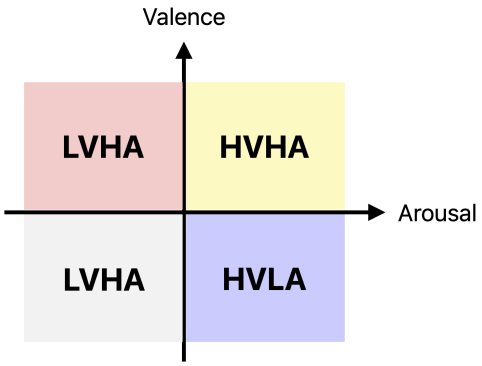


Fig. 4: 4-Quadrants Russell's circumplex model.

Russell's circumplex model of affect [12] represents emotional states on a two-dimensional plane defined by *valence*—the degree of pleasantness or positivity—and *arousal*—the level of activation or energy. It is widely adopted in music analysis since aligns well with musical experience, where variations in tempo, mode, rhythm, and intensity correlate strongly with shifts in arousal and valence. The EMOPIA dataset [13] adopted in this study exploits a 4-quadrants version of the Russell model (Fig. 4): quadrant Q1 High Valence–High Arousal (HVHA, e.g., joy or excitement), quadrant Q2 Low Valence–High Arousal (LVHA, e.g., tension or anger), quadrant Q3 Low Valence–Low Arousal (LVLA, e.g., sadness or calm), and quadrant Q4 High Valence–Low Arousal (HVLA, e.g., serenity or relaxation).

5.2 The EMOPIA Dataset

EMOPIA [13] is an emotion-labeled dataset for symbolic-domain music analysis and generation. It contains a total of 1087 music clips from 387 songs distributed along the four classes: 250 clips for Q1 (HVHA), 265 clips for Q2 (LVHA), 253 clips for Q3 (LVLA), 310 clips for Q4 (HVLA).

Due to copyright restrictions, music pieces in the dataset are delivered in the MIDI form. In order to perform audio analysis, we rendered the MIDI files using two high-quality virtual piano instruments, within the `fluidsynth` [1] audio plugin: the *Salamander Grand Piano* and the *Yamaha Disklavier Pro*. The former is a Yamaha C5, recorded with two AKG C414 microphones disposed in an AB position, at circa 12 cm above the strings. Samples are recorded at 48 kHz 24 bit with 16 velocity layers. The latter is modeled through a collection of 1212 samples recorded at a sampling frequency of 44.1 kHz. These instruments were chosen for their high audio fidelity, wide dynamic range, and detailed velocity-layer sampling, as well as their unique acoustic properties and distinct timbre. After the rendering, each file was resampled to mono 16 kHz format to meet the requirements of the DistilHuBERT model.

[1] https://www.fluidsynth.org

Components
CPU Intel Core i7-6700K 4 GHz, 4 cores, 8 threads
GPU NVIDIA GeForce GTX 1080 Ti 12 GB GDDR5, 3584 CUDA cores, 1582 MHz

Table 1: Hardware specifications of the machine used in the experiments.

The use of two instruments was motivated by the need of increasing data diversity and enabling oversampling during training. The EMOPIA dataset is affected by class imbalance, with uneven distribution across emotional labels. To address this, we adopt an oversampling strategy that balances the dataset by augmenting minority classes: instead of merely duplicating samples, we generate additional data by reproducing the same MIDI files using a different piano instrument, introducing timbral and expressive variation. This results in two distinct audio versions of each piece. Additionally, we apply audio-specific data augmentation techniques, including *pitch shifting* (modifying pitch without altering tempo), *noise addition* (simulating background noise), and *reverb* (emulating different acoustic environments). Specifically, we applied three different augmentation scenarios: 1) pitch shifting by three semitones up; 2) random Gaussian noise component addition, scaled by a factor of 0.0005; 3) pitch shifting by two semitones up, random Gaussian noise component addition, scaled by a factor of 0.0005 and reverb addition of two delay lines, one by 0.04s and the other by 0.8s.

5.3 Experimental Setup and Model Calibration

All experiments were conducted on the same machine (characteristics reported in Table 1) to avoid possible

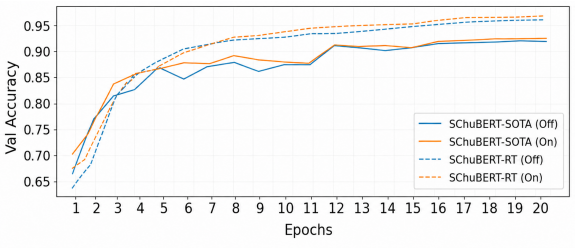


Fig. 5: Accuracy on the validation set with and without oversampling for SCHuBERT-SOTA.

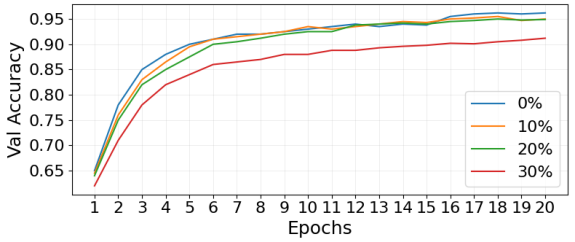
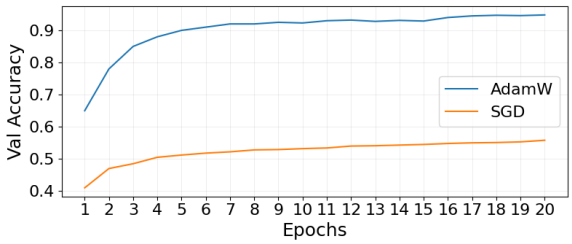
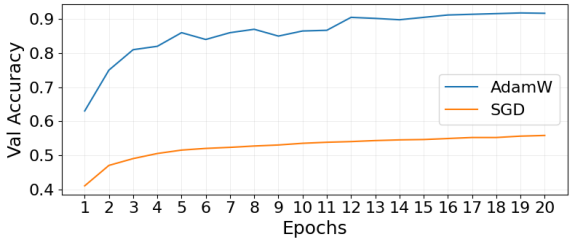


Fig. 6: Accuracy on the validation set at different augmentation probabilities for SCHuBERT-RT.



(a) SCHuBERT-RT



(b) SCHuBERT-SOTA

Fig. 7: Accuracy on the validation set using different optimization algorithms for both SCHuBERT-RT and SCHuBERT-SOTA.

fluctuations in performance between tests due to hardware changes, and considered both real-time and offline operation scenarios. The first scenario, which will be referred to as SCHuBERT-RT, was designed to assess the real-time processing capabilities of our proposed method. The second one, named SCHuBERT-SOTA, served as benchmark and did not exploit the real-time capabilities of our method.

The EMOPIA dataset was split as follows: 80% for training, 10% for validation, and 10% for testing. The validation portion was exploited for a model calibration procedure aimed at assessing the impact of *oversampling*, *data augmentation*, *batch size variation* and *optimization algorithm* on both models. The remaining model parameters adopted during the calibration stage (reported in Table 2, under the *Baseline* section) were kept fixed. The outcomes of the model calibration phase are summarized in the following.

Impact of oversampling. Figure 5 shows a slight positive impact of the dataset oversampling on the accuracy of SCHuBERT-SOTA. Conversely, oversampling brought no noticeable benefit to the accuracy of the SCHuBERT-RT model.

Impact of data augmentation. Figure 6 shows limited impact on the accuracy of SCHuBERT-RT when varying the probability ($pA = 0\%, 10\%, 20\%, 30\%$) of applying augmentation to the input data. We observed comparable results on SCHuBERT-SOTA (results not reported due to page limitations).

Choice of batch size. We tested the impact of different batch size values (16, 32, 64), i.e., the number of training samples processed before the model updates its parameters, on both models and no noticeable impact on the accuracy of either models was observed (results not reported due to space constraints).

Choice of optimization algorithm. Figure 7a and Figure 7b show a great improvement in accuracy when applying *AdamW* optimization techniques [27], in comparison to *Stochastic Gradient Descent (SGD)* [28], for both scenarios.

Based on the results of these preliminary experiments we selected the final parameters configuration (reported in Table 2, under the section *Optimized parameters*) to be used in all the subsequent experiments.

5.4 Results

We evaluated both models using standard classification metrics, namely confusion matrix, accuracy, precision,

	SCHuBERT-RT	SCHuBERT-SOTA
Baseline		
Window Size (s)	**0.5**	**3.0**
Hop Length (s)	**0.1**	**1.0**
Plateau Patience	2	2
Plateau Factor	0.5	0.5
Learning Rate (10^{-4})	10^{-4}	10^{-4}
Epochs	20	20
Optimized parameters		
Optimizer	AdamW	AdamW
Batch Size	32	32
Oversampling	**No**	**Yes**
Augmentation	No	No

Table 2: Baseline parameters used for model tuning and optimized parameters obtained in the tuning prcedure.

recall, and F1 score, across three complementary tasks designed to assess different aspects of emotional recognition. The first task, *4Q Classification*, consists of assigning each audio excerpt to one of the four quadrants of the valence–arousal (V–A) space, thus evaluating the model's ability to jointly capture both emotional dimensions. The second task, *Binary Arousal Classification*, focuses on discriminating between low and high arousal states, independently of valence, thereby assessing the model's sensitivity to variations in energy and intensity. The third task, *Binary Valence Classification*, evaluates the capability of distinguishing between negative and positive emotional content, regardless of arousal level. This multi-task evaluation protocol allows for a more comprehensive analysis of the model performance, enabling us to separately investigate the contribution of each emotional dimension, as well as their combined representation in the 4Q setting.

Confusion matrices for the *4Q Classification* task (Table 3) indicate that the SCHuBERT-RT model consistently outperforms SCHuBERT-SOTA across all emotional classes. The strong diagonal values highlight a high number of true positives, particularly in the LVHA quadrant, suggesting accurate identification of tense, negatively valenced excerpts. The limited off-diagonal values reveal only minor misclassifications, primarily occurring between different arousal levels. These observations are further supported by the results in Table 5, which report *accuracy*, *precision*, *recall*, and *F1-score*. Overall, the metrics confirm a stable performance advantage of SCHuBERT-RT, while SCHuBERT-SOTA shows consistently lower precision.

Similarly, the confusion matrices in Table 4 for the *Binary Arousal Classification* and *Binary Valence Classification* tasks demonstrate the superior capability of SCHuBERT-RT in distinguishing both high vs. low arousal and high vs. low valence, in comparison to SCHuBERT-SOTA. This improvement can be attributed to SCHuBERT-RT's shorter window and higher hop-length ratio. Additionally, both models exhibit better discrimination in arousal classification than in valence classification.

In order to assess the effectiveness of our approach, we also compare SCHuBERT with the start-of-art solutions on the task of *short chunk emotion recognition* [13], which consists of predicting the emotional content of audio excerpts with a length of 3 seconds. As reported in Table 6, SCHuBERT-SOTA, specifically tailored to perform audio-domain PMER in equal conditions as other related works on this task, significantly overcomes the performance of all other currently available models.

Finally, to assess the real-time performance of SCHuBERT-RT, we conducted a timing assessment by measuring both the *average pre-processing time* and the *average inference time*. As illustrated in Fig. 2, the pre-processing stage also includes MIDI-to-audio conversion in order to realistically simulate a real-time audio acquisition pipeline. For a comprehensive evaluation, 50,000 music clips (with some samples reused across iterations) were processed per iteration, for a total of 10 iterations. The experiments were carried out on both CPU and GPU systems, as detailed in Table 1. The results highlight a stable processing time across iterations on both CPU and GPU, indicating consistent performance of the system under repeated workloads. In particular, the GPU significantly outperforms the CPU, reducing the total processing time from approximately 20 ms to around 3 ms, while maintaining a negligible pre-processing overhead ($\approx$ 1 ms) in both cases.

6 Conclusion and Future work

In this work, we introduced SCHuBERT, a real-time end-to-end model for Piano Music Emotion Recognition and a protocol for multimodal dataset construction from live performances. Built upon the DistilHuBERT architecture for speech representation, SCHuBERT adapts self-supervised learning to music, processing raw waveforms to capture expressive emotional

(a) SCHuBERT-RT

T\P	HVHA	LVHA	LVLA	HVLA
HVHA	0.969	0.013	0.008	0.009
LVHA	0.008	0.981	0.007	0.004
LVLA	0.003	0.005	0.966	0.026
HVLA	0.003	0.001	0.016	0.980

(b) SOTA Model

T\P	HVHA	LVHA	LVLA	HVLA
HVHA	0.926	0.042	0.018	0.014
LVHA	0.058	0.912	0.017	0.013
LVLA	0.018	0.011	0.925	0.046
HVLA	0.009	0.006	0.053	0.932

Table 3: Confusion matrices for SCHuBERT-RT (a) and SCHuBERT-SOTA (b) models on the 4Q classification task.

(a) RT Arousal

T\P	LA	HA
LA	0.994	0.006
HA	0.014	0.986

(b) RT Valence

T\P	LV	HV
LV	0.980	0.020
HV	0.019	0.981

(c) SOTA Arousal

T\P	LA	HA
LA	0.978	0.022
HA	0.031	0.969

(d) SOTA Valence

T\P	LV	HV
LV	0.932	0.068
HV	0.059	0.941

Table 4: Confusion matrices of the for SCHuBERT-RT and SCHuBERT-SOTA models for the arousal (a, c) and valence (b, d) binary classification tasks.

Model Version	Accur.	Prec.	Recall	F1 Score
SCHuBERT-RT	0.975	0.975	0.975	0.975
SCHuBERT-SOTA	0.918	0.924	0.924	0.924

Table 5: Accuracy, precision, recall and F1 score of the SCHuBERT-RT and SCHuBERT-SOTA models on the 4Q classification task.

Model	4Q	A	V
SCHuBERT-SOTA	**0.918**	**0.973**	**0.936**
FF-D1-Mel-network [24]	0.865	0.912	0.743
Mel-network [24]	0.843	0.908	0.794
SCMA [22]	0.714	N/A	N/A
SAF Model [26]	0.692	0.884	0.869
Short-chunk ResNet [13]	0.667	0.887	0.704
Logistic regression	0.523	0.919	0.558

Table 6: Comparison between SCHuBERT-SOTA and state of the art methods in PMER.

cues without relying on handcrafted or symbolic features. On the EMOPIA dataset, SCHuBERT achieved state-of-the-art results in both four-quadrant and binary valence–arousal (V–A) tasks. Its real-time variant (SCHuBERT-RT) demonstrated comparable accuracy to the offline model while maintaining sub-20 ms CPU and sub-5 ms GPU latency, confirming its suitability for interactive scenarios. These results validate the effectiveness of self-supervised transformer architectures in combining high predictive performance with low-latency processing in affective music systems.

The proposed framework demonstrates strong potential for integration within Virtual Reality (VR) and immersive environments, as its real-time capabilities enable continuous affect-aware interaction between performers and virtual spaces, supporting applications such as adaptive sound rendering, emotion-driven virtual agents, and responsive musical environments.

Future work will focus on extending the emotional representation beyond the standard V–A model by incorporating finer-grained and higher-dimensional affective spaces enriched with a higher number of emotion categories, thus enabling the recognition of a broader range of expressive states. Moreover, we envision the integration of SCHuBERT into fully interactive VR systems for real-time multimodal affective feedback.

7 Acknowledgments

We acknowledge the support of the MUSMET project funded by the EIC Pathfinder Open scheme of the European Commission (grant agreement no. 101184379). This manuscript reflects only the authors' views and opinions and neither the European Union nor the European Innovation Council can be considered responsible for them.

References

[1] Canazza, S., Poli, G. D., Rodà, A., and Vidolin, A., "Expressiveness in Music Performance: Analysis, Models, Mapping, Encoding," 2013.

[2] Cancino-Chacón, C. E., Grachten, M., Goebl, W., and Widmer, G., "Computational Models of Expressive Music Performance: A Comprehensive and Critical Review," *Frontiers in Digital Humanities*, Volume 5 - 2018, 2018, ISSN 2297-2668, doi: 10.3389/fdigh.2018.00025.

[3] Jicol, C., Wan, C. H., Doling, B., Illingworth, C. H., Yoon, J., Headey, C., Lutteroth, C., Proulx, M. J., Petrini, K., and O'Neill, E., "Effects of Emotion and Agency on Presence in Virtual Reality," in *Proceedings of the 2021 CHI Conference on*

Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021, ISBN 9781450380966, doi:10.1145/3411764.3445588.

[4] Gall, D., Roth, D., Stauffert, J.-P., Zarges, J., and Latoschik, M. E., "Embodiment in Virtual Reality Intensifies Emotional Responses to Virtual Stimuli," *Frontiers in Psychology*, Volume 12 - 2021, 2021, ISSN 1664-1078, doi:10.3389/fpsyg.2021.674179.

[5] Yannakakis, G. N. and Togelius, J., "Experience-Driven Procedural Content Generation," *IEEE Transactions on Affective Computing*, 2(3), pp. 147–161, 2011, doi:10.1109/T-AFFC.2011.6.

[6] Forero, J., Mendes, M., and Bernardes, G., "Emotional Machines: Automatic Affective Virtual Environment Generation," 2023, doi:10.5281/zenodo.8329857.

[7] Wilson, R., "Aesthetic and Technical Strategies for Networked Music Performance," *AI and Society*, 38(5), pp. 1871–1884, 2023, doi:10.1007/s00146-020-01099-4.

[8] Turchet, L., "Musical Metaverse: vision, opportunities, and challenges," *Personal and Ubiquitous Computing*, 27(5), pp. 1811–1827, 2023.

[9] Turchet, L., Boem, A., Ayoub, O., Malandrino, F., Llorca, J., Fischione, C., and Rottondi, C., "Towards a Networking Architecture for the Musical Metaverse: The MUSMET Vision," in *2025 IEEE 6th International Symposium on the Internet of Sounds (IS2)*, pp. 1–8, IEEE, 2025.

[10] Chang, H.-J., Yang, S.-w., and Lee, H.-y., "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7087–7091, IEEE, 2022.

[11] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A., "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, 29, pp. 3451–3460, 2021.

[12] Russell, J. A., "Journal of Personality and Social Psychology," *Journal of Personality and Social Psychology*, 39(6), p. 1161–1178, 1980, doi:doi.org/10.1037/h0077714.

[13] Hung, H.-T., Ching, J., Doh, S., Kim, N., Nam, J., and Yang, Y.-H., "EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation," in *Proc. Int. Society for Music Information Retrieval Conf.*, 2021.

[14] Buccoli, M., Zanoni, M., Fazekas, G., Sarti, A., and Sandler, M. B., "A Higher-Dimensional Expansion of Affective Norms for English Terms for Music Tagging." in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, pp. 316–322, ISMIR, 2016, doi:10.5281/zenodo.1415194.

[15] Yang, Y.-H., Lin, Y.-C., Su, Y.-F., and Chen, H. H., "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, 16(2), pp. 448–457, 2008.

[16] Lu, L., Liu, D., and Zhang, H.-J., "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on audio, speech, and language processing*, 14(1), pp. 5–18, 2006.

[17] Weninger, F., Eyben, F., and Schuller, B., "On-line continuous-time music mood regression with deep recurrent neural networks," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5412–5416, IEEE, 2014.

[18] Li, X., Tian, J., Xu, M., Ning, Y., and Cai, L., "DBLSTM-based multi-scale fusion for dynamic emotion prediction in music," in *2016 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6, IEEE, 2016.

[19] Dong, Y., Yang, X., Zhao, X., and Li, J., "Bidirectional convolutional recurrent sparse network (BCRSN): an efficient model for music emotion recognition," *IEEE Transactions on Multimedia*, 21(12), pp. 3150–3163, 2019.

[20] Liu, H., Fang, Y., and Huang, Q., "Music emotion recognition using a variant of recurrent neural network," in *2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)*, pp. 15–18, Atlantis Press, 2019.

[21] Yang, P.-T., Kuang, S.-M., Wu, C.-C., and Hsu, J.-L., "Predicting music emotion by using convolutional neural network," in *International Conference on Human-Computer Interaction*, pp. 266–275, Springer, 2020.

[22] Xiao, Y., Ruan, H., Zhao, X., Jin, P., and Cai, X., "Music emotion recognition using multi-head self-attention-based models," in *International Conference on Intelligent Computing*, pp. 101–114, Springer, 2023.

[23] Chou, Y., Chen, I., Ching, J., Chang, C., and Yang, Y., "MidiBERT-Piano: Large-scale Pre-training for Symbolic Music Classification Tasks," *Journal of Creative Music Systems*, 8(1), 2024.

[24] Wang, H., Zhong, W., Ma, L., Ye, L., and Zhang, Q., "Emotional Quality Evaluation for Generated Music Based on Emotion Recognition Model," in *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pp. 1–6, 2022, doi:10.1109/ICMEW56448.2022.9859459.

[25] Zheng, L. and Li, C., "Real-Time Emotion-Based piano music generation using generative adversarial network (GAN)," *IEEE Access*, 12, pp. 87489–87500, 2024.

[26] Zhao, J. and Yoshii, K., "Multimodal multifaceted music emotion recognition based on self-attentive fusion of psychology-inspired symbolic and acoustic features," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1641–1645, IEEE, 2023.

[27] Loshchilov, I. and Hutter, F., "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2019.

[28] Robbins, H. and Monro, S., "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22(3), pp. 400–407, 1951.