Dissertation Title:

The Emotional Bridge: Translating Human Affect into Audiovisual Performance

Yuxin Chen

24012749

MSc Creative Computing in the Creative Industries (Modular)

University of the Arts London: Creative Computing Institute

Supervisor: Jessica Anderson

Submission Date: November 20, 2025

Chapter1

Introduction

1.1 Motivation

In many algorithmically driven creative systems, emotion has been operationalised as an output condition—a controllable result rather than an intrinsic structural process. Within affective music generation, control is often defined as the user's ability to specify a desired emotional state and to adjust musical parameters so that the system outputs music expressing that affect (Dash and Agres 2024). Similarly, Takahashi et al. (2022) propose an automatic harmonisation and tempo-arrangement framework that functions on given melodies under emotional constraints, treating affect as an external rule to shape musical form. This output-oriented framing also extends to human–computer interaction and audiovisual design. Moreover, another example, an emotion-aware interface analyses the tone of voice and facial expression to control communicative feedback (Rojas et al. 2022), while projects such as Breathing Scarf visualizes emotional arousal through bodily data translated into light and colour (Cochrane et al. 2022). Also, live performances use mappings between gesture, sound, and emotion to construct and maintain clear perceptibility and control (West et al. 2021). Similar strategies are used in visualization systems, where emotional cues are encoded as changes in colour, brightness, and motion (Nunes et al. 2024; Hu et al. 2025). Together, these practices show a shared logic of emotional externalisation: emotion becomes a measurable display, positioned at the endpoint of generation rather than dive into the compositional logic itself.

Recent research in affective computing and music information retrieval has highlighted the epistemic limits of this paradigm. Peintner et al. (2025)'s work shows that the most emotion-recognition models use simplified categories, such as the nine discrete emotions from the GEMS model. It is mostly because of the limits to tools or data. However, the way of thinking leaves out how emotion can change over time or relate to other things. By contrast, Chin et al. (2024) show a different idea. They point out that the changes in harmony and tone can affect people's emotion which means that emotion comes from how music is built, not just it sounds on the surface. Likewise, Hu et al. (2025) also support this, they show that sound and colour can match by emotion when a system learns from both at once. Their work shows that emotion has structure and can connect across different senses.

Those works give me a lot of inspiration. My research follows this new direction. I want to treat emotion not as a surface effect or a output, but as something that helps shape both sound and visuals as they change over time. I don't think emotion as something added at

the end, instead the system I'm building uses emotion as a core part of how things are made.

In my system, emotion will work as a "control button", which affects how rhythm moves, how chords change, and how visuals react. It doesn't just show emotion, also creates through it. Through this way, emotion helps drive how sound and visual grow and change together.

1.2 Research Goals

My research will explore how emotion can act as a creative tool in real-time audiovisual work. Rather than using emotion as a descriptive label or end result, I prefer to explore how emotion can shape rhythm, chords, and visuals as they happen. My goal is to build a system where emotion takes an active role. It helps guide how things are made, changed and interact with each other in real time.

Through this process, I want to bulid a method that bridge emotional flow with how structure is formed. My system will create original works based on emotional input.

Chapter2

Background

2.1 Conceptual and Theoretical Context

Emotion has always played a key role in art and design. In ancient Greece, the idea of catharsis showed how strong feelings were part of the experience. Later, during the Baroque period, the "Doctrine of Affections" also linked artistic form to emotional expression (Langer 1953). John Dewey (1934) later described art as an organic process of experience, where meaning arises from the continuous exchange between body, emotion, and material. In this sense, affect is not merely the content of art but one of its generative logics.

Within contemporary artificial-intelligence and human–computer-interaction contexts, emotion has increasingly been quantified by algorithms, reduced to a controllable variable or a classification label. As McCarthy and Wright (2004) note, technological systems often functionalise emotional experience, shifting the human–machine relation from embodied empathy toward parameter control. Höök (2018) therefore proposes that body and emotion should be treated as the ontology of design rather than passive data inputs. This philosophical foundation guides the present research: reinserting emotional experience into algorithmic generation so that affect regains agency in shaping musical

and visual structures.

Against this background, recent developments in interactive art and performance have sought to restore emotion's active role through embodied and participatory systems. In the development of interactive art and performance research, affect has gradually come to be understood as a dynamic element within interaction rather than a signal to be recognised and displayed. Building on somaesthetic and affect-as-aesthetic perspectives that view bodily feeling as central to how interaction takes shape (Höök, 2009), recent systems for live performance explore real-time co-creation in which gesture, voice, and computational models shape musical expressivity during play rather than merely at the output stage (Borovik et al., 2023). Within HCI, new work explicitly frames real-time tracking of participants' affective states as an interactive modality for performance and installation, integrating sensing and mapping into the generative loop of the piece (Hosale et al., 2024). Taken together, these studies motivate my approach: I treat real-time emotion recognition and data mapping as a generative dialogue between music and image, aiming to make the system a co-creator of evolving form rather than a mimic that applies emotion post hoc.

2.2 Technical Context of Emotion-Driven Creation

My system captures the performer's facial expressions in real time using a fine-tuned model based on the FER2013 dataset, a public dataset containing 28,709 training and 3,589 test grayscale 48×48 images categorized into seven emotions (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral) (FER2013 Dataset, 2013).

Real-time detection employed the YOLOv8 architecture, a one-stage model capable of performing facial localisation and emotion classification within a single network. It was used because it runs fast and stays stable when working with live video. Other models like ResNet-18 need extra steps, such as cropping the face and then running classification. These steps make things slower when processing each frame.

2.2.1 Acoustic Feature Extraction

In computer-based music analysis, low-level descriptors (LLDs) are the basic way to represent sound. They give clear, trackable values that show how sound changes over time.

LLDs come from the MPEG-7 standard. They measure short-term traits like pitch, sharpness, harmonic quality and timbre. These values are taken from short parts of the audio signal and are saved as single numbers or short lists. This setup makes it possible to study small changes in loudness, tone and rhythm. Many emotion-based music systems use this kind of data.

Modern tools like librosa make it easier to do this work. They include built-in ways to pull out and study these sound features quickly (McFee et al., 2015). These tools help connect creative ideas with data, letting systems match sound behaviour with emotional states in real time.

Based on this method, a piece of music can be described using six main features: RMS energy, spectral centroid, spectral rollof, zero-crossing rate, tempo and event density. Each one shows a different part of the sound. Together, they help describe how loud, bright or rhythmically active a sound is.

The Root-Mean-Square (RMS) value expresses the mean power of the waveform over time and correlates strongly with perceived loudness or energetic intensity(音频特征提取_音频的特征提取-CSDN 博客, 2023).

$$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{n=1}^{N} x[n]^2}$$

The Spectral Centroid defines the “centre of mass” of the spectrum, indicating where most spectral energy is concentrated and corresponding perceptually to brightness or sharpness.(Spectral centroid, 2025)

$$\mathrm{centroid} = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)},$$

The Spectral Rolloff marks the frequency below which a fixed proportion (r, commonly 0.85) of total spectral energy lies(Librosa.feature.spectral_rolloff, 2025).

The Zero-Crossing Rate (ZCR) counts how often the waveform changes sign within a frame(过零率, 2022):

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} \mathbb{I}\{s_t s_{t-1} < 0\}$$

Finally, Event Density (or Onset Rate) measures the number of detected sound onsets per second.

When normalised and weighted, these six acoustic features together make an energy score that reflects how strong and active the sound feels.

Studies in music psychology show that these features are closely linked to the arousal

level of emotion. Schubert (2004) found that loudness and tempo are the main factors that predict arousal. They have a much stronger effect than other features such as pitch or timbre, while their effect on valence (pleasantness) is much weaker. This means that the feeling of energy in sound is strongly connected to how excited or calm people feel.

Ilie and Thompson (2006) also tested intensity, tempo, and pitch in both music and speech. Their study showed that when both intensity and tempo go up, people feel more aroused in each case. However, pitch works differently. It changes how pleasant something fells in music and in speech in different ways. These findings show that the energy in sound can help predict how emotionally active it feels.

2.2.2 Tonal and Harmonic Analysis

Except volume and energy, the way tones are arranged also affects how music shows emotion.

The Constant-Q Transform (CQT) changes a sound signal into a format that follows the musical scale. Each step matches a semitone. Because the frequency bins grow in a logarithmic way, the same pattern repeats across octaves. This makes it easier to compare notes with different pitches (Wikipedia, 2025). Then, we can take the average of CQT values across octaves to get chroma features. These show pitch classes like C, D or E are most active, no matter how high or low they sound.

By comparing these pitch patterns with stored key templates, algorithms can guess the most likely key or mode of the music. This follows known models of how we hear tonal structure (Krumhansl, 2000).

2.2.3 Affective Dimensions and Emotional Mapping

The Circumplex Model of Affect (Russell, 1980) explains emotion as a two-dimensional space with valence (from pleasant to unpleasant) and arousal (from calm to excited) as its main axes. Unlike fixed emotion categories, this model treats feeling as a continuous change rather than a set of labels.

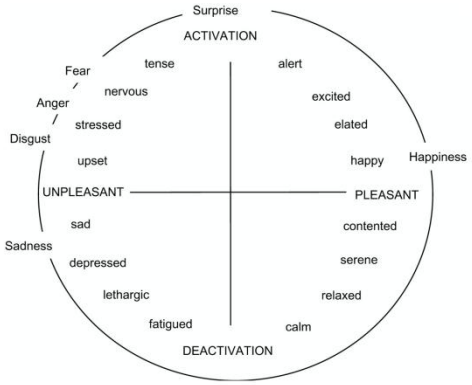

figure 2.1 Russell's (1980) Circumplex Models

In studies of affective computing and music psychology, researchers have found that arousal is closely linked to sound intensity, tempo, and brightness (Duman et al., 2022), while valence relates more to musical mode (major or minor) and timbre warmth (Carraturo et al., 2024). This framework makes it possible to translate measurable sound and visual features into emotional values that can guide generative systems in real time.

2.2.4 Real-Time Interaction Protocols

For computational systems that combine sensing and generative output, low-latency communication is essential.

Open Sound Control (OSC) improves on MIDI by allowing high-precision, time-stamped messages to be sent over local or network connections(Derivative, 2024). Its flexible system lets users send many kinds of data like emotion and space, between tools such as Python, Touchdesigner and MaxMsp. (Wright, Freed & Momeni, 2017).

These kinds of systems shape performace by using feedback. The technology acts like a partner that responds during the creative process (Tanaka, 1970).

2.3 Artistic and Conceptual Precedents

2.3.1 Emotion and Sound

Nowadays, many artists are now turning emotions into sound. Some recent interactive works track people's facial expressions or body movements, and link them to sound settings. This lets people help create music just by showing their emotion (X Ciaowei et al., 2023).

Another system makes sound based on emotion in real time. It changes volume, tone and rhythm depending on how the performer feels (Hariyady et al., 2024).

A well-known example is Heart Space by Krista Kim (2024–2025). This work turns heartbeats and breathing into sound and light. The changing body signals create a live mix of rhythms and tones. It shows that emotion is not still, it moves and flows in a shared sound and light space.

figure2.2 Heart Space (Krista Kim, 2025)

2.3.2 Emotion, Vision, and Embodiment

A similar shift is happening in visual and live media. Many new art installations now use sensors to turn body and emotion signals into changing visuals.

For example, Rajapakse (2025) uses EEG brainwave data to create images with diffusion models. The changing mental and emotional states become a stream of moving visuals, showing thought as a shifting landscape. In Transhuman Ansambl (Ivsic et al., 2024)'s work, virtual performers react to the audience's voice. This creates a live exchange between people and machine voices.

Emotion shows the main force behind the interaction instead of the end result in these works. By using sensors with generative tools, artists link inner feeling with outer form, letting emotion guide what we see and hear.

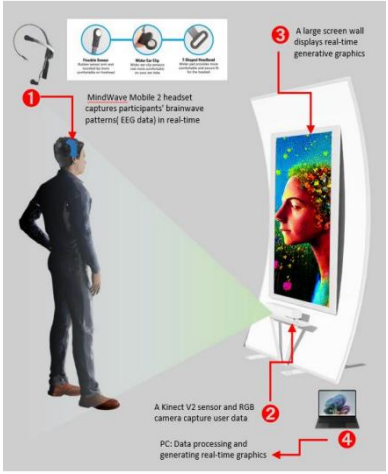

figure2.3 ThoughtDiffusion (Rajapakse, 2025)

Chapter3

Methodology

My research uses practice-oriented method, which include machine learning, sound analysis and visual interaction. In the execution, I devide the experiment into four main parts, and construct the system step by step. In this process, my core goal always beexploring how emotional data guide the production process.

3.1 Emotion Recognition

Facial emotion detection is the first step I take, it is also the basic of the whole project. I choose the FER2013 dataset to train my emotion model, and fine-tuned the original dataset, which can help improve the accuracy of detection. The results are sent as control data, and the timed detection mode is used instead of the continuous detection modes.

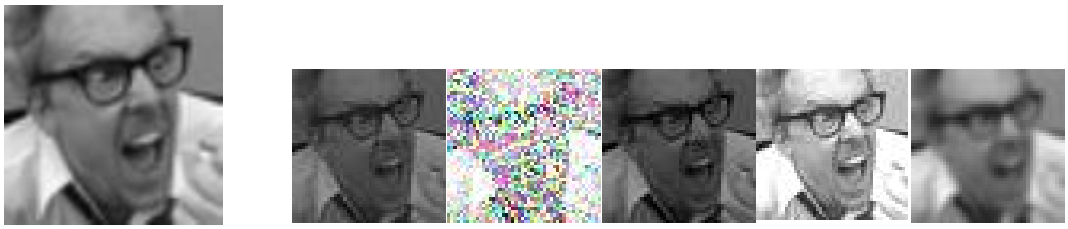

figure 3.1 fer2013 (fine-tuned)

3.2 Music Analysis and Mapping

In music analysis part, I use two datasets to study the way of music construction and the changes of energy. Berlin dataset includes 255 short loops, and Techno learn includes complete techno music, which used to observe the overall structure and tonal flow.

3.3 Music Generation System

In the music generation component, the system operates as a real-time loop-based architecture. Each time an emotion is sampled at the end of a section, the target chord and the energy intensity for the next musical block are determined together. These values drive the selection of harmonic, percussive, and ornamental loops from the Berlin Underground library.

Using these parameters, the system assembles loops into multi-layer rhythmic modules. Tonal material is pitch-shifted to match the predicted chord, while drum and percussive layers are selected according to the estimated energy level. To maintain continuity

between blocks, the system uses an end-of-block sampling gate to stabilise the emotional input, and a preloading mechanism that renders the next block several seconds earlier. A short crossfade is applied at the block boundary, ensuring smooth transitions and preventing audible discontinuities during real-time performance.

3.4 Visual Generation and Integration

The visual part has two main stages: the creation of start frame and end frame, and real-time control. Firstly, using text-to-image mode in runway to create image of keyframes, then put these two frames into video mode and turn into video clips. Lastly, import the clips into Touchdesigner to add some effects. This process ensures that the sound and visuals are unified through the experience.

Chapter4

Results

4.1 Emotion Recognition

At the beginning of the project, I tried to interpret emotions through body movements rather than facial expressions. Therefore, I used the keyword “people dancing in the dance floor”, collected and organized 89 YouTube videos to construct a dataset based on movements, but the accuracy of the results was insufficient. To solve this problem, I turned my attention to capturing the performer’s (single) facial expressions. Research has shown that audiences often sense the emotion conveyed by the performers.

As mentioned in Section 3.1, I fine-tuned the FER2013 dataset. I added various data such as blurring, reduce brightness and dark spots to images to simulate the lighting of live performance. The final training set included 114,836 images, and the validation set included 28,709 images, covering seven emotion categories. The model has a resolution of 224*224 pixels and is trained in batches of 64 images for 120 rounds. It achieves a Top-1 accuracy of 98.8% and a Top-5 accuracy of 100% on validation set.

In practical use, the model continuously monitors the facial expressions of single performer’s face, and calculates the average result within a short time widow at the end of each section. This method can effectively reduce the recognition errors caused by the subtle facial tremors.

4.2 Music Feature Learning

To prepare the sound layer, I build two datasets:

1.Berlin Underground – 255 open-source techno loops (Bass, Drum, Kick, Pad, Synth, Claps, Hats, Percussion, Rides, SFX, Snares, Toms).

2. Techno Learn – ten full techno tracks around 130 BPM for studying long-term energy and harmonic flow.

Berlin Underground: Loop Analysis

When I analysed the full Berlin Underground library, I divided everything into two groups: tonal loops and non-tonal loops. Only the loops that contain a stable pitch structure—bass, pad, synth, chord, keys, lead, arp, and melody—are treated as tonal. These loops go through a chroma-CQT check to estimate their possible root note, and the results are saved in a dedicated tonal-only metadata file.

Everything else—kick, clap, snare, hat, drum, top, ride, perc, sfx, and so on—falls into the non-tonal group. These sounds focus more on rhythm, noise, or texture, and don't carry any reliable harmonic information, so they are not assigned a key. Non-tonal loops shape the rhythm and energy, while the tonal loops handle the chord direction and take part in key-based shifting later in the system.

This separation keeps the whole engine stable: tonal layers stay harmonically correct, and the rhythmic layers remain free, flexible, and punchy. It lets the system keep its musical structure without limiting the character of the drums and textures.

path	type	rms	centroid	rolloff	zcr	tempo	density	key_guess
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 5.wav	Kick	0.265071	375.69	485.39	0.006429	0.0	3.6667	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 11.wav	Kick	0.184373	786.65	1454.84	0.01534	0.0	11.0002	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 10.wav	Kick	0.158239	779.88	2093.66	0.004578	0.0	1.8334	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 4.wav	Kick	0.312513	200.34	351.71	0.004517	0.0	5.5001	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 6.wav	Kick	0.168004	893.87	2029.06	0.028483	0.0	3.6667	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 12.wav	Kick	0.151618	2286.67	3991.89	0.1891	0.0	6.4165	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 13.wav	Kick	0.01898	3486.99	6193.77	0.284336	0.0	2.7498	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 7.wav	Kick	0.220792	492.82	793.14	0.005432	0.0	3.6667	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 3.wav	Kick	0.157005	2077.43	4548.44	0.082865	0.0	7.3335	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 17.wav	Kick	0.122026	2013.64	4221.85	0.072367	0.0	3.6667	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 16.wav	Kick	0.268819	226.57	389.84	0.005371	0.0	5.5001	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 2.wav	Kick	0.229733	503.25	625.81	0.006144	0.0	3.6667	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 14.wav	Kick	0.170943	1412.3	2846.87	0.087646	0.0	3.6667	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 15.wav	Kick	0.194176	287.51	455.79	0.005412	0.0	3.6667	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 1.wav	Kick	0.307472	963.91	2147.94	0.009521	0.0	7.3476	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 9.wav	Kick	0.206144	938.81	2066.29	0.02004	0.0	7.3335	
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Kick/Kick 8.wav	Kick	0.234896	305.39	453.54	0.005941	0.0	3.6667	

/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_003_110_Am.wav	Pad	0.054039	2007.01	4001.91	0.094134	0.0	4.0104	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_012_110_Am.wav	Pad	0.029666	1415.51	2570.1	0.082114	0.0	2.6927	B
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_006_110_Am.wav	Pad	0.183788	2330.21	4393.0	0.119542	0.0	6.7604	C
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_001_110_Am.wav	Pad	0.133463	882.21	1442.41	0.064182	0.0	1.4896	C
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_010_110_Am.wav	Pad	0.051671	2780.1	4431.46	0.193191	0.0	1.1458	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_015_110_Am.wav	Pad	0.171187	2792.29	5132.63	0.1429	0.0	4.9271	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_009_110_Am.wav	Pad	0.139527	1193.92	2670.89	0.056889	0.0	4.6979	E
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_004_110_Am.wav	Pad	0.202884	1417.66	2496.53	0.060346	0.0	2.1771	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_016_110_Am.wav	Pad	0.097004	811.5	1369.65	0.043373	0.0	7.1041	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_007_110_Am.wav	Pad	0.208631	1433.99	2658.23	0.050401	0.0	5.6146	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_002_110_Am.wav	Pad	0.112685	1346.11	1966.71	0.0869	0.0	6.0729	C
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_013_110_Am.wav	Pad	0.192856	1806.2	4163.25	0.052552	0.0	4.8698	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_014_110_Am.wav	Pad	0.165914	828.31	1381.09	0.05455	0.0	5.3854	C
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_008_110_Am.wav	Pad	0.103035	1533.2	3071.9	0.072586	0.0	8.5365	A#
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_005_110_Am.wav	Pad	0.122368	759.43	1271.5	0.052725	0.0	3.9531	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/Pad loop/Padloop_011_110_Am.wav	Pad	0.109029	1300.03	1803.95	0.11579	0.0	7.3333	G
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/bass loop/ Bassloop_003_110_Am.wav	Bass	0.250105	965.5	2084.91	0.01288	0.0	7.3333	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/bass loop/ Bassloop_012_110_Am.wav	Bass	0.272218	1226.5	2625.88	0.025996	0.0	7.3333	A
/Volumes/TOSHIBAEXT/ualproject/dataset/techno dataset/Berlin underground/bass loop/ Bassloop_006_110_Am.wav	Bass	0.209248	628.88	1237.93	0.009673	0.0	3.4375	A

table 4.1 loop_meta_v3_tonal_only.csv (cropped section)

I also use chroma-CQT profiles to estimate the key of each loop. I save the results in a file called loop_meta_v2.csv, which includes the file path, loop type, guessed key, and normalised energy. This key data is later used to match loops to chords.

Techno Learn: Full-Track Analysis

Across the ten full tracks in the Techno Learn dataset, each piece was analysed and divided into short sections of about fifteen seconds. The energy – brightness curves show that the tracks do not move in one straight line from low to high. Instead, they rise and fall in repeating waves. Low-energy parts (around 0.2 – 0.35) appear not only at the beginning but also return midway as short breaks. Build and middle sections stay around 0.4 – 0.5, and the peak sections usually remain between 0.5 and 0.6 rather than spiking sharply. Overall, the tracks follow a looping rhythm of building up, releasing, and building up again.

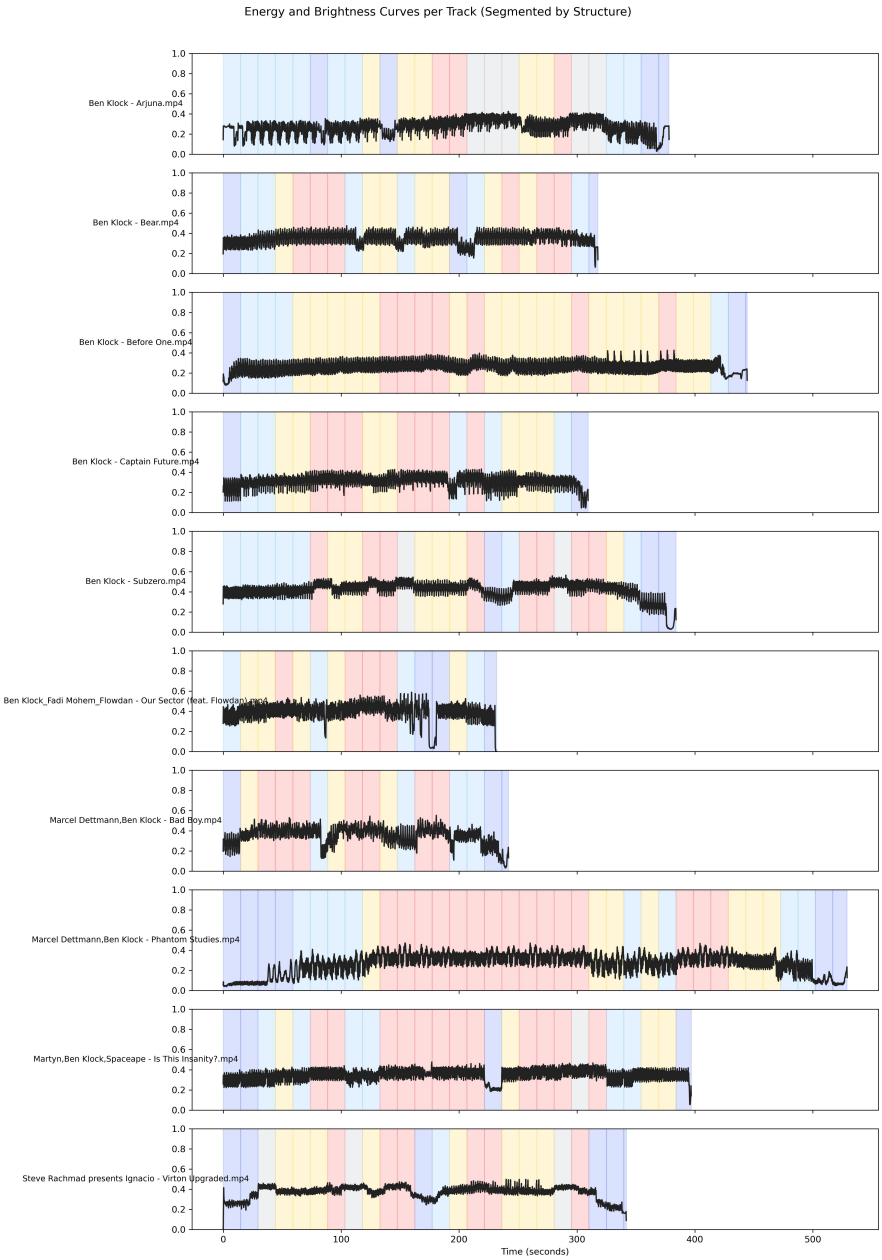

table 4.2 Full-track analysis

Based on these labelled sections, the data was further turned into a simple layer-activation template covering seven common elements in Techno: Kick, Bass, Pad, Synth, Hats, Snare/Clap and Perc. On average, intro sections keep only pads and light hats, builds add bass and some percussion, middles stabilise with more layers, peaks open almost all layers, and breakdowns reduce most drum elements while keeping the harmonic parts. Together, these two outputs describe a stable pattern in how intensity and texture change across the tracks, providing a straightforward structural reference for the later real-time performance system.

4.3 Emotion-Driven Music Generation

The final audio engine connects facial-expression recognition to a library of pre-composed loop materials, forming a real-time generative music system built around an 8-bar structure. At 130 BPM, each section lasts around 14.8 seconds.

When a section approaches its ending, the system opens a short 0.8-second sampling window to stabilise the emotional input. During this window, several predictions are collected and merged into a single label through a majority-vote process. This label is then mapped to a target chord and an energy level for the next block.

According to these parameters, the engine selects at least one rhythmic layer (Kick, Clap, Percussion) and one tonal layer (Bass, Pad, Synth) from the Berlin Underground dataset. Tonal loops are pitch-shifted to match the chosen chord, while rhythmic and ornamental layers are filtered based on the expected intensity.

The next block is assembled about two seconds in advance. A 1.2-second crossfade is applied at the section boundary so that transitions remain smooth and stable during real-time performance.

4.4 Visual Generation and Integration

For the visual aspect of system, I hope to merge the music's emotion into visuals and movements. At the beginning, I tried to use ComfyUI to generate real-time looping videos based on emotions. What I thought was generate first and last frame images, and useing the images to generate looping videos. However, my computer could not handle it. The diffusion model ran poorly, it took over five minutes for each image.

Therefore, I turn to use offline approach: Runway, a model that can be used on website. I create start and end frame, and use those frames generate video clips. Due to the limitation of Runway's video time, I edited the clips to ensure them can cover the music's timing. Then, I imported them into Touchdesigner.

figure 4.2 An example of start and end frames

In Touchdesigner, I use Switch Top to connect video clips which corresponding different

emotions. A Python script tells system what current emotion is through OSC signal, then selecting which video to play. When model recognizes new emotions, OSC triggers video switching and updating parameters, maintaining harmony between music, visuals and emotions.

In order to enable visuals change following emotion and sounds in real-time, I use Transform Top and Displace Top to add a responsive visual effect. The energy of mid-range audio frequenciesis are converted into MIDI signals and then mapped to Displace weight in real-time. The stronger the sound energy, the more pronounced the visual movements.

figure 4.3 Sound visualization effect

The visual design of my project inspired by Anyma's style: a visual language that combines accuracy of mechanical control with emotion expression. The core of visual is a "mechanical lady" who express emotion as a human. She exists between cool algorithms and human's senses.

This character shows the main concept of the project: People and machine can build a collaborated relationship during the created process. In this framework, emotion becomes a bridge that connect organic beings and digital world, creating a connection between them in audiovisual interaction.

figure4.3 Anyma, Y do I - The End Of Genesys [Official Visualizer]

Chapter5

Final Works

At this stage, my system's design goal is to perform the emotion be a primary driver. It tries to use emotions as force instead of relying on manual input to control sound and visual changes, to make the audiovisual flow more like human inner states.

The final work "Emotion Driven Performance" combines facial emotion recognition, chord control and real-time visual feedback into an interconnected system. In this system, performer's facial expressions keep influencing music and visual elements, resulting the entire performance shows as a dynamic evolution driven by emotion.

5.1 Emotion-to-Sound Translation

The performance begins from emotion input. The fine-tuned YOLOv8-FER2013 model keeps reading the performer's expressions and categorises them into seven basic emotions. These categories are not only recognition results, but also act as the control signals that shape how the music behaves.

Each emotion drives a different musical tendency. "Happy" often brings major chords such as C or G, brighter synthesiser tones and slightly stronger rhythms. "Sad" usually leads to minor keys like A or E, softer pads and fewer percussive layers. "Surprise" introduces more unstable or tension-based loops, while "angry" emphasises lower frequencies and rougher textures.

With the system updating emotion once every 8-bar loop, the music is able to shift according to the performer's changing mood without producing abrupt changes. The result is a real-time interaction where the music feels as if it is responding to the performer's inner rhythm, forming a more direct link between facial expression and sound.

5.2 Emotion-to-Visual Translation

The same emotion signal also drives the variation of visual layer. I pre-generate a 16s narrative video for each emotion in Runway. Then, I import them into Touchdesigner. When the system receive new emotion lable through OSC, the Swich Top shifts corresaponding video contents automatically.

To avoid static play, the system transforms the audio characteristics especially mid

frequency energy to MIDI values, controlling displace weight dynamically. The resulting visual warping beats to the rhythm of sound, creating a sense of breath.

The core of visual is a “mechanical lady”, which symbolizing the tension between machine senses and human emotion. I design a “dual-body” effect for her: Through overlaying two offset versions, to make her shows a subtle inconsistencies. This represents the uncertainty and delay in the process of the system “understand” the emotion.

5.3 Emotion-Driven Feedback Loop

In this system, emotion recognition, sound control and visual changes become a interactive cycle. The entire process is driven by emotion: It first influence the generation of music, then sounds influence visual changes. Visual feedback then turns to affect audiences or performer’s sense of emotion, this feedback re-enters to the system, continue influencing following sounds and visuals.

This system does not use emotion as static set of data, but treat it as a continously changing dynamic factor. Emotion becomes a starting point and connecting point of a feedback loop, which making people’s subjective experience and machine’s response to flow in the sound and visual. They co-create the rhythm and atmosphere by using a detectable and sensible way.

Chapter 6

Evaluation

6.1 Technical Evaluation

6.1.1 Emotion Detection

The fine-tuned YOLOv8-FER2013 model shows a high accuracy on the validation set, reaching 98.8%. The model can stay relatively stable recognition even when lighting conditions changed.

This stability enables the system keep receiving emotion data instead of frequent manual actions, providing a smoother input source for music and visual control. The reliable transfer of emotion information enhances the entire system's expressiveness and coherence in live performance.

6.1.2 Emotion-Driven Music System

The system samples emotion only at the end of each music section. As an 8-bar block approaches its ending, it opens a 0.8-second detection window and collects several emotion predictions. The most frequent result is then used as the label for the next block. This method reduces abrupt fluctuations caused by single-frame misclassifications and helps keep the musical transitions controlled.

The emotion–chord mapping maintains a stable harmonic centre while still allowing the music to shift gradually with the performer's expression. Each emotion corresponds to a predefined chord pool and an energy value, enabling adjustments in musical density and intensity without disturbing overall tonal continuity.

During testing, the delay between the detected emotion and the resulting musical change remained below 200 ms. This level of responsiveness allows the interaction to feel direct while avoiding overly reactive behaviour that could destabilise the musical flow.

Overall, the system's behaviour suggests that a small amount of temporal smoothing at the decision stage can improve the stability of emotion-driven music transitions, while still preserving a sense of responsiveness during live performance.

6.1.3 Emotion-Driven Visual System

The Python audio engine and Touchdesigner achieves close time synchronization through OSC. There is almost no delay during entire operation, ensuring the excellent real-time interaction effect.

Visual responds dynamically to changes in the energy of mid level band. When energy is low, visuals remain smooth and subtle. When energy increases, visuals exhibit more obvious displacment and transformation. This presents a dynamic, emotion-related visual rhythm.

figure 6.1 Effect Comparison

6.2 Aesthetic Evaluation

The core of this work is it transforms relationship between performer, systems and the stage. Emotion is no longer an element in music or visuals, it becomes a bridge connecting sounds and visuals, playing an important role in the whole performance.

Facial expressions are more than signals to be recognized, they become a form of communication in this system, bringing performer's feelings to the visual control system. Whenever performer's face change, the chords and visuals, even the rhythm shifts. Emotion in this case acts like a thread, binding the auditory and visuals together. The sounds become something you can see, and visuals become something you can hear. The human face is no longer a tool for expressing feelings, it becomes a starting point of the whole performance.

This approach brings a new form of performance. It breaks the wall between DJs and VJs. Emotions become their shared language, allowing both sounds and visuals to flow with performer's mood. Systems, people, music and visuals are not separated, they interact in the same rhythm.

Through this design, the work is exploring a question: Can emotion "perform" without any langauge? In this system, emotion is not only an observed object, but also the force that push everything.

6.3 Limitations and Future Work

During testing, the system showed several notable limitations. Firstly, the real-time performance and the accuracy of facial emotion recognition is not always stable. For instance, it always misunderstands “neutral” as “sad”, and “disgust” rarely appears. This may be attributed to the lack of disgust samples in FER2013 dataset. Due to the facial expression itself has multi-dimensional nature, relying on facial inputs to control system presents limitations. In future versions, it is possible to consider introducing additional inputs such as body posture, heart rate variation or movement intensity, to construct a more comprehensive emotional map.

In the music section, each piece’s generation involves selection of new rhythm tracks, including fundamental drums like kick drums. This structure may cause breaks or delay. A more desiralbe approach is keep the rhythmic consistency while changing the chords and harmonic structure. It can enhance the whole performance’s sense of flow effectively.

The visual part is limited by hardware, and cannot yet achieve a fully real-time generation. The current approach is pre-generate video clips in Runway for each emotion, then import them to Touchdesigner for displacement and transformation through OSC. This approach is suitable for testing control logic, but it is still a semi-realtime workflow. The next step is to construct a system which can generate and modify the visual output in real-time, and fully tied to the audio engine. To achieve those functions, the requirements for computing resources are quite high. Not only needs faster model response, but also needs multiple GPUs to collaborate to generate high-quality frames for different emotions. These challenges bring a deeper thought: In generative art, emotion expressions not only rely on creator’s thoughts and feelings, but also depend on technical resource. In other words, the “emotion” that the audience eventually senses sometimes depend on the computational power that powers these softwares.

Chapter 7

Discussion

7.1 Technical Contribution

In practice, the end-of-block sampling produced a more stable emotional input. By collecting several predictions within a short window and selecting the most frequent result, the system avoided abrupt changes caused by single-frame fluctuations. This contributed to smoother transitions between musical sections.

The chord and energy mapping maintained a consistent tonal direction while still allowing emotional differences to be reflected through variations in texture, density, and timbre. The pre-building of the next block and the 1.2-second crossfade also reduced the abruptness that can occur in loop-based generation, helping the output join more continuously from one section to the next.

During testing, the delay between detecting an emotion and hearing its effect stayed below 200 ms. This level of responsiveness supported a clear connection between facial expression and system output, while still keeping the overall musical structure stable.

7.2 Aesthetic Contribution

Visually and conceptually, I want to explore a question: When emotions guided by machines, how would it reshape the relationship between performer, sounds and visuals?

Turning emotions into signal which can be “seen” and “measured”, so that they no longer the feelings hidden in the performer’s heart, but become something that can be recognized, adjusted and even participate as a part of live performance actively. Every emotion detection, no matter if it is accurate, can influence sounds and visuals, becoming a part of performance. Those changes let us understand a fact: There are always some differences between the human understanding of the emotion and the machine calculation of the emotion.

The work around “control” and “release”expand. The performer’s face is the starting point triggered by system. And the following changes in the visuals and sounds are generated by the system based on its own logic. This kind of feedback is not fully controllable. Because of this uncertainty, the performance becomes more open and spontaneous.

My inspiration partly comes from Anyma's performace. I'm trying to construct a visual space. The "mechanial lady" has emotions as a human, responding based on performer's facial movements. To express the uncertainty of emotions, I designed a visual effect that shows two overlapping versions of the same robot. It looks both familiar and strange. These fuzzy and bias represent data and real feelings do not always match perfectly. Because of these imperfections, it shows a unusual and poetic expression.

In the process of creation, the system is no longer as a tool to be manipulated. It has become more like a partner with personality and judgment. Sometimes it's accurate, but sometimes it responds something different. It participates the performance by its own way, completing emotional expression with human. In this interaction, a new collaboration emerged between human and machine.

7.3 Closing Remarks

The original aim of my project was testing the ability of the computer to process emotions. However, in the process of experiment, the point of research gradually turned into a more fundamental problem: When we sharing the control with computers, what does that mean? I realized that if emotions are turned into data, they no longer exist as their original and subjective form. Instead, they responded and perceiced by neutral networks, becoming signals flowing between code, system loops and sounds. This process subtly changes the nature of the emotion.

Sometimes there are some errors or rhythm disruptions, it doesn't mean failure. By contrast, to some extent, the errors construct a way that human communicate with machine: a collision between human expression and algorithm logic. This uncertainty becomes a part of creative process.

From a longer perspective, this system remind us that emotions can not only become the source of data, but also can be a part of design logic. The real-time emotional feedback mechanism offers the possibility of building a more direct and more emotional interaction. And some new technology such as diffusion model, multiple systems, may promote the development of visual, auditory and emotion dimensions.

Of course, building such a system still faces several technical challenges. It relies on high-performance hardware, including multiple GPUs, high-speed processor, and emotion detection and response tools designed for live environments.

Therefore, this work is not only a technical experiment in the field of digital art, but also push as to think about broader questions: If we hope that machines can respond to human's emotions, how much time, computing power and creativity would this process require? How does it influence the system's mode of operation and performer's expression and identity when emotions are transformed, recreated or even reconstructed?

Chapter 8

Conclusion

This project tried to discuss how emotion being a bridge of human expression and machine output in real-time audio-visual environment. Through integrate facial emotion detection, mapping the emotion on music and visual control technology based on OSC, I have initially developed a system that can adjust sound and visual in real time according to the emotion changes, gave emotions an opportunity to play a certain role in guiding the creative process.

The focus of the project was not on achieving the accuarcy and stability of the system, but on exploring the function of emotion in the performance. When there is a delay or an error, these faults often become a unique part of the performance. These phenomenan suggest that there are certain differences in the way machines and humans understand emtions, and those differences may have practical significance.

In the process of experiments, I gradually realize that when emotion are transformed into data, their expressiveness may increase, but at the same time, they also exhibit instability and uncertainty.

In conclusion, this project provide a perspective, to consider whether emotions can not only serve as an input of generation system, but also potentially influence the performance structure and execution sequence. It attempts to transform emotions into an interactive mechanism, exploring the role that emotions play in human-machine collaboration: Machines may not be passive responders to emotions, but may also be involved in the process of emotion generation.

Bibliography

(No date) *音频特征提取_音频的特征提取-CSDN 博客*. Available at: https://blog.csdn.net/qq_30129009/article/details/129624036 (Accessed: 10 November 2025).

Anyma (2025) *Anyma, Y do I - The End Of Genesys [Official Visualizer]*, *YouTube*. Available at: https://www.youtube.com/watch?v=Ciq7MRYEIyM (Accessed: 06 November 2025).

Borovik, I. and Viro, V. (2024) *Real-time co-creation of expressive music performances using speech and gestures*, *Zenodo*. Available at: https://zenodo.org/records/11189321 (Accessed: 05 November 2025).

Camurri, A. *et al.* (2004) 'Multimodal analysis of expressive gesture in music and dance performances', *Lecture Notes in Computer Science*, pp. 20–39. doi:10.1007/978-3-540-24598-8_3.

Carraturo, G. *et al.* (2025) 'The major-minor mode dichotomy in music perception', *Physics of Life Reviews*, 52, pp. 80–106. doi:10.1016/j.plrev.2024.11.017.

Chin, J.D., Clark, M. and Doryab, A. (2024) 'Creating emotion-evoking music to communicate wellness for users with diverse musical backgrounds', *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 418–424. doi:10.1145/3675094.3678495.

Cochrane, K. *et al.* (2022) 'Breathing scarf: Using a first-person research method to design a wearable for emotional regulation', *Sixteenth International Conference on Tangible, Embedded, and Embodied Interaction*, pp. 1–19. doi:10.1145/3490149.3501330.

Constant-Q transform (2025) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Constant-Q_transform (Accessed: 05 November 2025).

Dalibor Mitrovic, D., Zeppelzauer, M. and Eidenberger, H. (2006) 'Analysis of the Data Quality of Audio Features of Environmental Sounds', *Journal of Universal Knowledge Management*, 1(1), pp. 4–17.

Dash, A. and Agres, K. (2024) ‘AI-based Affective Music Generation Systems: A review of methods and challenges’, *ACM Computing Surveys*, 56(11), pp. 1–34. doi:10.1145/3672554.

Derivative (2021) *OSC in DAT - touchdesigner documentation*, *Derivative*. Available at: https://docs.derivative.ca/OSC_In_DAT (Accessed: 06 November 2025).

Duman, D. *et al.* (2022) ‘Music we move to: Spotify audio features and reasons for listening’, *PLOS ONE*, 17(9). doi:10.1371/journal.pone.0275228.

Eerola, T. and Vuoskoski, J.K. (2012) ‘A review of Music and Emotion Studies: Approaches, emotion models, and stimuli’, *Music Perception*, 30(3), pp. 307–340. doi:10.1525/mp.2012.30.3.307.

Elshamy1, M.R. and ECE, 1Klipsch School of (2024) *P-yolov8: Efficient and accurate real-time detection of distracted driving ††thanks: The authors would like to thank the following funding agencies: NSF grant 2219680.*, *P-YOLOv8: Efficient and Accurate Real-Time Detection of Distracted Driving The authors would like to thank the following funding agencies: NSF grant 2219680.* Available at: https://arxiv.org/html/2410.15602v1 (Accessed: 05 November 2025).

Hariyady, H. *et al.* (2024) ‘Harmonizing emotion and sound: A novel framework for procedural sound generation based on emotional dynamics’, *JOIV: International Journal on Informatics Visualization*, 8(4), p. 2479. doi:10.62527/joiv.8.4.3101.

Hosale, M.-D. *et al.* (2024) ‘Beam workshop: Biophysical expression, affect & movement’, *Proceedings of the 9th International Conference on Movement and Computing*, pp. 1–3. doi:10.1145/3658852.3661318.

Hö ok, K. (2018) *Designing with the body: Somaesthetic interaction design*. Cambridge, MA: The MIT Press.

Https://london.mocomuseum.com/krista-kim-heartspace/ (no date) *Krista Kim: Heart Space | Moco Museum London Exhibition*. Available at: https://london.mocomuseum.com/krista-kim-heartspace/ (Accessed: 06 November 2025).

Hu, J. *et al.* (2025) ‘Music2Palette: Emotion-aligned color palette generation via Cross-Modal Representation Learning’, *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6615–6624. doi:10.1145/3746027.3754921.

Höök, K. (2009) ‘Affective loop experiences: Designing for interactional embodiment’, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535), pp. 3585–3595. doi:10.1098/rstb.2009.0202.

Ilie, G. and Thompson, W.F. (2006) ‘A comparison of acoustic cues in music and speech for three dimensions of affect’, *Music Perception*, 23(4), pp. 319–330. doi:10.1525/mp.2006.23.4.319.

Ivsic, L., McCormack, J. and Dziekan, V. (2024) *Transhuman ANSAMBL - voice beyond language*, *arXiv.org*. Available at: https://doi.org/10.48550/arXiv.2405.03134 (Accessed: 06 November 2025).

Krumhansl, C.L. (2000) ‘Rhythm and pitch in music cognition.’, *Psychological Bulletin*, 126(1), pp. 159–179. doi:10.1037/0033-2909.126.1.159.

Librosa.feature.spectral_rolloff (no date) *librosa.feature.spectral_rolloff - librosa 0.11.0 documentation*. Available at: https://librosa.org/doc/latest/generated/librosa.feature.spectral_rolloff.html (Accessed: 10 November 2025).

McFee, B. *et al.* (2015) ‘Librosa: Audio and Music Signal Analysis in python’, *Proceedings of the Python in Science Conference*, pp. 18–24. doi:10.25080/majora-7b98e3ed-003.

Nunes, C. *et al.* (2024) ‘Thunder: A design process to build emotionally engaging music visualizations’, *Proceedings of the XXIII Brazilian Symposium on Human Factors in Computing Systems*, pp. 1–15. doi:10.1145/3702038.3702077.

Peintner, A. *et al.* (2025) ‘Nuanced music emotion recognition via a semi-supervised multi-relational Graph Neural Network’, *Transactions of the International Society for Music Information Retrieval*, 8(1), pp. 140–153. doi:10.5334/tismir.235.

Pertical (2023) *Pertical/Yolov8: Yolov8 Ь in PyTorch > ONNX > CoreML > TFLite*, *GitHub*. Available at: https://github.com/Pertical/YOLOv8 (Accessed: 05 November 2025).

Rajapakse, R.P.C.J. (2025) ‘Thoughtdiffusion: An interactive installation for exploring neuro-art from EEG data with stable diffusion models’, *Proceedings of International Conference on Artificial Life and Robotics*, 30, pp. 662–667. doi:10.5954/icarob.2025.os24-2.

Rojas, C. *et al.* (2022) ‘Towards enhancing empathy through emotion augmented remote communication’, *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–9. doi:10.1145/3491101.3519797.

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

Sambare, M. (2020) *Fer-2013*, *Kaggle*. Available at: https://www.kaggle.com/datasets/msambare/fer2013 (Accessed: 05 November 2025).

Schubert, E. (2004) ‘Modeling perceived emotion with continuous musical features’, *Music Perception*, 21(4), pp. 561–585. doi:10.1525/mp.2004.21.4.561.

Spectral centroid (2025) *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Spectral_centroid (Accessed: 10 November 2025).

Takahashi, T. and Barthet, M. (no date) *Emotion-driven harmonisation and tempo arrangement of melodies using transfer learning*, *ISMIR 2022: Schedule*. Available at: https://ismir2022program.ismir.net/poster_80.html (Accessed: 22 November 2025).

Tanaka, A. (1970) *Mapping out instruments, affordances, and mobiles*, *Mapping Out Instruments, Affordances, and Mobiles - Goldsmiths*. Available at: https://research.gold.ac.uk/id/eprint/6834 (Accessed: 06 November 2025).

Tresnawati, D., Nurhidayanti, S. and Lestari, N. (2025) ‘A comparison of yolov8 series performance in student facial expressions detection on online learning’, *Jurnal Online Informatika*, 10(1), pp. 93–104. doi:10.15575/join.v10i1.1390.

West, T. *et al.* (2021) ‘Making mappings: Design criteria for live performance’, *NIME 2021* [Preprint]. doi:10.21428/92fbeb44.04f0fc35.

Wright, M., Freed, A. and Momeni, A. (2017) ‘2003: OpenSound Control: State of the art 2003’, *Current Research in Systematic Musicology*, pp. 125–145. doi:10.1007/978-3-319-47214-0_9.

Wright, P. and McCarthy, J. (2004) *Technology as experience*. MIT Press.

Xiaowei, C. and Zainuddin, I. (2023) *Exploring emotional responses in interactive installation art: A convergence of affective computing and artificial intelligence* [Preprint]. doi:10.20944/preprints202307.1792.v1.

Yarwood, M. (no date) *Russell's (1980) Circumplex Models*, *PennState*. Available at: https://psu.pb.unizin.org/psych425/chapter/circumplex-models/.

过零率 (2022) *Wikipedia*. Available at: https://zh.wikipedia.org/wiki/%E8%BF%87%E9%9B%B6%E7%8E%87 (Accessed: 10 November 2025).