Proceedings of CONF-CDS 2026 Symposium: Computer Vision-Based Multimodal Cognitive Load Estimation for Adaptive Media Communication
DOI: 10.54254/2755-2721/2026.GL36283

Can You See Music? A Study of Whether a Hand-Drawn Visualization Strengthens the Emotion of AI-Generated Music

Judy Jiaqi Liu

United World College South East Asia, East Campus, Singapore, Singapore
jiaqijudyliu@gmail.com

Abstract. AI tools can now write a piece of music for a target emotion, but the sound by itself does not always make that emotion clear to a listener. We built a system called Can You See Music that plays AI-generated music and, at the same time, draws a simple hand-drawn animation that moves with the sound. We asked two questions. First, does adding the drawing make the emotion feel stronger or more enjoyable? Second, which emotions does the system show well, and which ones does it show poorly? We ran a within-subjects study with 12 high-school students. Each student rated six emotion clips in two conditions: music only, and music with the drawing. Adding the drawing raised the average intensity rating by a small amount, from 4.32 to 4.51 on a 7-point scale, but this difference was not statistically reliable ($t(11) = 0.83$, $p = .43$, $dz = 0.24$). Enjoyment did not change (4.28 vs 4.22). When we measured how well each emotion matched its drawing, calm and sad scored high and were rated consistently, while fearful scored lowest and split the group: some students felt it matched very well, and others felt it did not match at all. The open comments agreed with these numbers. Our study is small and exploratory, but it points to a clear next step. The drawings for high-energy negative emotions, such as fear and anger, are the weakest part of the system and should be redesigned first.

Keywords: Affective Computing, Music and Emotion, Audio Visualization, User Study, Valence and Arousal

1. Introduction

Music makes people feel things. A slow song can feel sad, and a fast song can feel exciting. Today, AI tools can take a target emotion and write a new song for it in a few seconds, and people are starting to use them to make music together [1]. This is useful for games, videos, and apps that want the right mood on demand. But there is a problem. When you hear an AI song with no other information, you cannot always tell which emotion it was meant to carry. The same clip might sound"tense" to one person and "sad" to another.

One way to help is to add something to see, not just something to hear. People often understand a feeling faster when a picture supports the sound, because the senses are linked in regular ways [2]. A movie scene feels scarier with the right visuals, even though the music is the same. So we asked a

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GL36283

simple question: if we show a drawing that moves with an AI song, does the emotion come through more strongly?

To test this, we built Can You See Music. The system plays a short AI-generated clip for one of six emotions and draws a hand-drawn animation that reacts to the sound in real time. We then ran a small study where students rated the clips with and without the drawing.

RQ1. Does adding the hand-drawn visualization make the emotion feel stronger (higher intensity) and more enjoyable than audio alone?

RQ2. Across the six emotions, which ones does the visualization match well, and which ones does it match poorly?

This paper makes three small contributions. First, we describe a working system that pairs AI music with a live hand-drawn visualization. Second, we report an exploratory user study with 12 students that measures intensity, enjoyment, and match. Third, we give a clear and honest picture of where the system works and where it does not, so that the next version can be improved on purpose rather than by guessing.

2. Background and related work

2.1. Describing emotion with two numbers

To study emotion, we first need a way to write it down. We use the circumplex model from Russell [3]. In this model, any feeling can be placed on two axes. Valence says how positive or negative the feeling is, from sad to happy. Arousal says how calm or energetic it is, from sleepy to excited. For example, "calm" has positive valence but low arousal, while "angry" has negative valence but high arousal. Research on music shows that these two numbers do a good job of capturing the emotion of a song [4]. We chose our six emotions so that they sit in different corners of this space. This lets us check not just one emotion, but a spread across the whole map.

2.2. Music and emotion

Researchers have long shown that music carries emotion through features such as tempo, key, and loudness [5]. Fast and loud often reads as high arousal, while slow and soft often reads as low arousal. AI music tools build on this, but they hide the controls. The user asks for a mood and gets a finished song. Our work sits on top of these tools and asks a follow-up question: once the song exists, can a picture help the listener feel the intended mood?

2.3. Showing sound

Turning sound into pictures is an old idea, from classic music visualizers to modern generative art. The senses do not connect at random. People match sounds to colors and shapes in regular ways, an effect known as crossmodal correspondence [2], and many of these music-to-color links are guided by the emotion in between [6]. Some people even see color when they hear music [7], and color, music, and feeling line up in lab studies [8]. Children also draw music in surprisingly consistent ways [9], and recent systems show that a well-planned visualization can carry a song's emotion to a viewer [10]. Most visualizers aim to look cool. Our goal is different. We want the drawing to support a specific emotion, not just to react to the beat.

DOI: 10.54254/2755-2721/2026.GL36283

3. The Can You See Music system

The system has two parts: the music and the drawing.

Music. We used an AI music tool to generate one short clip for each of six target emotions: happy, sad, calm, excited, angry, and fearful. Each emotion was given a target point in valence-arousal space before generation, shown in Table 1. The clips were kept short so that a student could rate many of them in one sitting.

Drawing. While a clip plays, the system runs a live visualizer. It reads the audio signal, measures features such as loudness and frequency content using a Fast Fourier Transform, and uses those features to drive a hand-drawn animation on screen. Louder and busier parts of the song make the drawing move more, and the color and shapes are chosen to fit the target emotion. The drawing is redrawn many times per second so that it stays in time with the music. We chose a rough, hand-drawn look on purpose, because a sketch can feel more human and less like a machine readout, and rough sketches are a known way to keep a design feeling open and friendly [11].

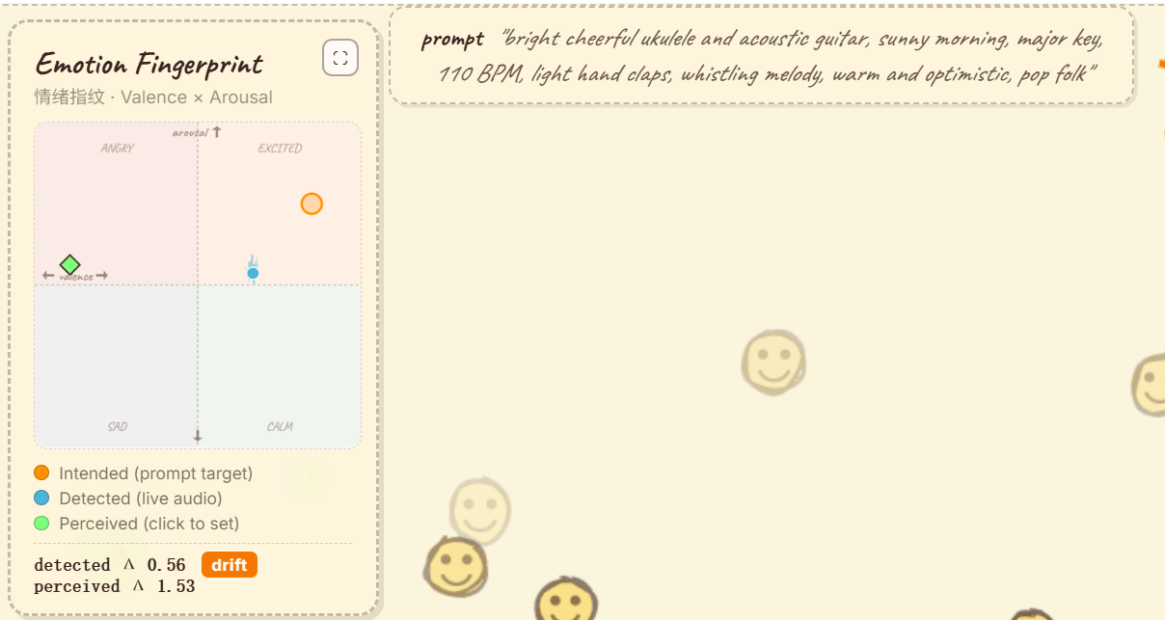

Figure 1. The Can You See Music system running the "happy" clip

The Emotion Fingerprint panel (top left) plots the intended, detected, and perceived emotion in valence-arousal space, while the live hand-drawn visualization reacts to the audio.

Table 1. The six emotions and their target valence (V) and arousal (A) values, from −1 to +1

Emotion	Valence	Arousal	Short description
Happy	+0.7	+0.5	Positive, fairly energetic
Excited	+0.6	+0.9	Positive, very energetic
Calm	+0.5	−0.7	Positive, low energy
Sad	−0.6	−0.5	Negative, low energy
Angry	−0.7	+0.8	Negative, high energy
Fearful	−0.8	+0.6	Negative, high energy

DOI: 10.54254/2755-2721/2026.GL36283

4. Method

4.1. Design

We used a within-subjects design with two independent variables, following standard practice for user studies in human-computer interaction [12]. The first variable was condition: A, music only, and B, music plus the drawing. The second was emotion, with six levels. Each student saw both conditions, so every person acted as their own comparison. To control for order effects, we split the trials into two blocks, one all-A and one all-B, and we balanced the starting block across students. Six students started with A then B, and six started with B then A.

4.2. Participants

Twelve high-school students took part (mean age 15.1 years, range 15 to 16). Six reported as male, four as female, one preferred not to say, and one did not report gender. For music background, three had formal training, seven had some experience, and two had none. Participants were recruited by convenience from our classmates and friends, and all gave consent before starting.

4.3. Materials and measures

The materials were the six AI clips and the visualizer, run in a web browser. After each clip, students answered short questions on a 7-point scale. Intensity asked how strong the emotion felt (1 = very weak, 7 = very strong). Enjoyment asked how much they enjoyed the experience. In condition B only, match asked how well the drawing fit the music, because this question only makes sense when there is a drawing to judge. At the end, students answered three open questions about the most memorable clip, the least matching clip, and any suggestions.

4.4. Procedure

Each session ran on one device and took about ten minutes. A student first read the consent screen and filled in age, gender, and music background. They then completed the first block of six trials, one per emotion, rating each clip right after hearing or seeing it. They completed the second block in the other condition, and finished with the open questions. The median time spent on a single trial was about 23 seconds.

4.5. Analysis

We collected 144 trials in total, 12 students by 12 trials each. We first computed descriptive statistics, the mean (M) and standard deviation (SD), for each condition and emotion. For RQ1 we ran paired t-tests comparing A and B on intensity (H1) and enjoyment (H2), and we report the effect size Cohen's dz next to each p-value. For RQ2 we compared match scores across the six emotions with a repeated-measures ANOVA (H3). Because our sample is small, we treat all tests as exploratory and we read effect sizes and patterns, not just p-values.

DOI: 10.54254/2755-2721/2026.GL36283

5. Results

5.1. Data preparation

No participants were removed. One file was saved in Apple Numbers format and was converted to CSV before merging. Two students happened to enter the same ID, "P07"; we kept them as two separate people because their timestamps and answers differed. We checked for careless answering, such as giving the same score to everything, and found none. Most scores sat between 4 and 5.5 on the 7-point scale, so the scale was not hitting its ceiling or floor.

5.2. H1: does the drawing make the emotion stronger

Adding the drawing raised average intensity from M = 4.32 (SD = 0.65) in condition A to M = 4.51 (SD = 0.84) in condition B, a difference of +0.19. This change was in the predicted direction but was not statistically reliable, $t(11) = 0.83$, $p = .43$, $dz = 0.24$. Looking person by person, six students went up and six went down, so the average rise came mostly from two students who responded strongly to the drawings.

5.3. H2: does the drawing make it more enjoyable

Enjoyment did not change. It was M = 4.28 (SD = 0.73) for music only and M = 4.22 (SD = 0.78) for music plus drawing, a difference of −0.06, $t(11) = -0.33$, $p = .75$, $dz = -0.09$. This is a clear null result. The drawing may make the feeling slightly stronger, but it did not make the experience more pleasant.

5.4. H3: which emotions match best

Table 2 shows the average match score for each emotion in condition B, from highest to lowest. Sad scored highest and was the most consistent. Fearful scored lowest, and more importantly it had by far the largest spread (SD = 2.21). Its twelve ratings were 7, 6, 6, 2, 7, 7, 3, 5, 2, 6, 6, 1. Half the group felt the fearful drawing fit very well, and the other half felt it did not fit at all. A repeated-measures ANOVA across the six emotions was not significant, $F(5, 55) = 0.74$, $p = .60$, $\eta p2 = .06$, which means the average scores were similar. The real story is not in the averages but in the spread: calm and sad were a steady success, while fearful was an average that hides a deep disagreement.

Table 2. Match scores by emotion (condition B only), 7-point scale, N = 12

Emotion	Mean match	SD	Pattern
Sad	5.75	0.87	High and consistent
Happy	5.50	1.38	High
Calm	5.08	1.44	Solid
Excited	5.00	1.60	Mixed
Angry	5.00	1.35	Mixed
Fearful	4.83	2.21	Split in two

DOI: 10.54254/2755-2721/2026.GL36283

5.5. What students said

The open comments matched the numbers. Several students pointed to the high-energy clips as the weak spots, which is exactly where the match scores were lowest or most divided.

"The very intense song had a lot of fire shooting out. On its own it gave me a feeling of rescue and tension, like a chase."

P01, on the least matching clip

"The clip with the eyes and the low humming music felt mismatched."

P06, on the least matching clip

"The sad songs felt the strongest. The energetic song with the firework pieces was the least matching."

P07

Two students gave the system a direction to improve in. P021691 wanted"more sad songs,"the emotion that already worked best, and noted that one angry clip"had nothing angry"in it. P60 felt the strongest clips came from the strongest target emotions and suggested the drawings"can be made for stronger impacts."

6. Discussion

This section takes the numbers from Section 5 and says what they mean. We answer each research question in turn, then we step back and talk about what the results say for AI music in general, and what we would change in the design.

6.1. RQ1: does the drawing help

The short answer is "a little, and only in one way." Adding the drawing pushed the felt intensity up by a small amount, but the change was not reliable in our sample, and enjoyment did not move at all. So the drawing may make the emotion feel a bit more present, but it does not make the experience more fun.

The small rise in intensity fits the idea that the senses support each other. Because sound and image are linked in regular ways [2], and because color and music meet through emotion [6], a matching picture can add a second channel that points at the same feeling. That is the direction our intensity result took, even if it was weak. The flat enjoyment score is also worth a closer look. Past work on adding sound to a task found a gap between what people expect to help and what actually helps [13]. We may be seeing the same gap. A drawing feels like it should make things nicer, but it does not always do so. One simple reason is that strong is not the same as pleasant. A scary or angry drawing can pull you in without being nice to look at, so intensity can go up while enjoyment stays flat.

6.2. RQ2: which emotions work

The clearest result is about the type of emotion, not its average score. Low-arousal feelings, calm and sad, were matched well and almost everyone agreed. High-arousal negative feelings, angry and fearful, were the weakest, and fearful split the group in two. So the system is good at quiet emotions and shaky at loud, dark ones.

This pattern matches what the music research would predict. Quiet, low-arousal music tends to have clear and steady cues, such as a slow tempo and a soft sound [5], and these are easier to turn

DOI: 10.54254/2755-2721/2026.GL36283

into a steady picture. The valence and arousal model also captures these emotions well [4]. The split on fearful is the key design signal. Earlier work shows that people can share consistent links between music and images [9], and that a planned visualization can carry emotion across senses [10]. Our fearful drawing did not reach that shared reading. When half the group rates a match a 6 or 7 and the other half rates it a 1 or 2, the drawing is most likely leaning on one specific image, like fire or eyes, that some people connect to fear and others do not. The fix is not to make it stronger, but to make it more shared.

6.3. What this means for AI music

Our study sits on top of AI-generated music, so it is worth asking what the results say about that tool. AI music is now easy to make and is even used as a way for people to create together [1]. But a generated clip carries a target emotion that the listener cannot see. The label lives in the prompt, not in the sound. Our weakest cases show this clearly. One trained participant said an "angry" clip "had nothing angry" in it, which means the AI's intended emotion and the listener's reading did not match. This is the gap our visualization is trying to close.

There is a second point about AI and how people judge it. Research shows that just knowing something is "AI-made" can shift how people rate it, apart from the content itself [14]. In our study the music was AI-made for every clip and every condition, so this bias was held constant and cannot explain the differences we found between emotions. But it is a reason to be careful in future work, and a reason to give AI music some extra support, like a matching picture, so its intended emotion can be read more directly. In short, AI can produce the sound quickly, but making the intended emotion clear to a human is a separate problem, and that is the problem our system targets.

6.4. Design implications

Good design comes from matching the system to how people actually read it [15], and from treating each result as feedback for the next version [16]. Read this way, our data gives a clear to-do list. The quiet emotions are working and should be left alone. The loud, negative emotions need new drawings that more people can agree on. Because the problem is disagreement and not weakness, the next version should test drawings on several people before shipping, and keep the ones where the spread is small.

7. Limitation

• Small, narrow sample. We had only 12 students, almost all age 15, recruited by convenience. The results may not hold for other ages or groups.

• Exploratory tests. With this few people, our tests have low power. A small but real effect can easily miss significance, so we read effects and patterns, not p-values alone.

• Single session and few clips. Each emotion had one clip, so a "weak" emotion might reflect that one clip, not the emotion in general.

• Intensity is not one thing. "Strong" and "enjoyable" may be different feelings that our questions did not fully separate.

8. Future work

• Redesign the hardest emotions first. Fearful and angry are the weakest and least agreed-upon, so they should be the first to be redrawn and retested.

DOI: 10.54254/2755-2721/2026.GL36283

• Run a larger study. Our intensity effect was small (dz = 0.24). A future study with roughly 40 or more people could test whether that small rise is real.

• Separate engagement from enjoyment. We would ask about "how pulled in did you feel" and "how much did you like it" as two questions, since H2 suggests they come apart.

• Use more than one clip per emotion so that the result describes the emotion, not a single song.

9. Conclusion

We built a system that pairs AI-generated music with a live hand-drawn visualization, and we tested it with 12 students. Adding the drawing raised the felt intensity of the emotion by a small, non-significant amount and did not change enjoyment. The drawing worked best for quiet emotions like calm and sad, and worst for loud, negative emotions like fearful, where students disagreed sharply about whether it fit. The study is small and the findings are exploratory, but they are honest and they point somewhere useful. To make the system better, the next step is not to add more features, but to fix the few drawings that are clearly failing.

References

[1] Suh, M., Youngblom, E., Terry, M., & Cai, C. J. (2021). AI as social glue: Uncovering the roles of deep generative AI during social music composition. *CHI 2021*.

[2] Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics, 73*(4), 971-995.

[3] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178.

[4] Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music, 39*(1), 18-49.

[5] Gomez, P., & Danuser, B. (2007). Relationships between musical structure and psychophysiological measures of emotion. *Emotion, 7*(2), 377-387.

[6] Palmer, S. E., et al. (2013). Music-color associations are mediated by emotion. *PNAS, 110*(22), 8836-8841.

[7] Ward, J. (2013). Synesthesia. *Annual Review of Psychology, 64*, 49-75.

[8] Whiteford, K. L., Schloss, K. B., Helwig, N. E., & Palmer, S. E. (2018). Color, music, and emotion: Bach to the blues. *i-Perception, 9*(5), 1-25.

[9] Liu, S., Bu, J., Ye, H., Chen, J., Jiang, S., Tao, M., Guo, L., Wang, C., & Li, C. (2024). DoodleTunes: Interactive visual analysis of music-inspired children doodles with automated feature annotation. *CHI 2024*.

[10] Nunes, C., Reinbold, I., Castro, M., & Darin, T. (2024). Thunder: A design process to build emotionally engaging music visualizations. *IHC 2024*.

[11] Buxton, B. (2007). *Sketching User Experiences*. Morgan Kaufmann.

[12] Lazar, J., et al. (2017). *Research Methods in Human-Computer Interaction* (2nd ed.). Morgan Kaufmann.

[13] Newbold, J. W., et al. (2017). Using nature-based soundscapes to support task performance and mood. *CHI EA 2017*.

[14] Ragot, M., Martin, N., & Cojean, S. (2020). AI-generated vs. human artworks. A perception bias towards artificial intelligence? *CHI EA 2020*.

[15] Norman, D. (2013). *The Design of Everyday Things (Rev. ed.)*. Basic Books.

[16] Carroll, J. M. (Ed.). (2003). *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*. Morgan Kaufmann.